

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ



ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

School of Information Sciences and Technology

Department of Informatics

Athens, Greece

Master Thesis

in

Data Science

Chain of Thought Prompting for Intent Classification using Large Language Models

Dimitrios Koutsianos

Supervisor: Ion Androutsopoulos

Department of Informatics

Athens University of Economics and Business

Company Themos Stafylakis

Supervisors: Omilia Conversational Technology

Panagiotis Tassias

Omilia Conversational Technology

March 2024

Dimitrios Koutsianos

Chain of Thought Prompting for Intent Classification using Large Language Models

March 2024

Supervisor: Ion Androutsopoulos

Athens University of Economics and Business

School of Information Sciences and Technology

Department of Informatics, Information Processing Laboratory, Natural Language Processing Group

Athens, Greece

Abstract

Large Language Models (LLMs) like ChatGPT, GPT-4, and Llama have shown tremendous capacity in text understanding and generation. Among the most prominent emergent abilities of LLMs is the so-called Chain of Thought (**CoT**) prompting. The latter consists of a series of intermediate reasoning steps that can significantly improve the ability of LLMs to perform complex reasoning. Moreover, it provides an interpretable window into the behavior of the model, suggesting how it might have arrived at a particular answer and providing opportunities to debug where the reasoning path went wrong.

The goal of the suggested project is to examine the application of CoT prompting in sentence classification, and more specifically in intent classification. Estimating the intent of the user of a service is one of the core problems in task-oriented dialog systems. Traditionally, it has been approached as a classification problem requiring many training examples per intent and older classifiers are trained on a pre-defined set of intents. Such approaches are in general not scalable and hard to maintain, as they require one model per client due to the set of intents being client-dependent. An additional weakness of these approaches is that they do not reveal the way the model classifies phrases into intent classes. Explainability methods such as LIME or SHAP can be applied, however, they do not always yield intuitive explanations, as they merely estimate the contribution of each word of the phrase to the classifier’s decision, something that Chain of Thought Prompting may provide a solution for.

LLMs have exhibited strong performance in text classification, both in a 0-shot and a few-shot context, but not on the same level as older models that are designed specifically for classification. To address this performance gap, in this work a set of 5 candidate intents are added to the prompt. These intents are extracted utilizing a BERT-based classifier. To further help with the reduction of the performance gap, a description for each of the 5 candidate intents is also included in the prompt a description, which was created with the help of ChatGPT.

As for the Chain of Thought part of this thesis, we first tested a 0-shot method utilizing three phrases that generate Chain of Thought, *“Let’s Think Step by Step”*, *“Let’s Take a Deep Breath and work on this Step by Step”*, and a novel phrase, *“Show your Thoughts”*. We will show that in both datasets that are used, namely *CLINC-150* and *BANKING77*, the method that was outlined shows promising results.

Περίληψη

Μεγάλα γλωσσικά μοντέλα (LLM) όπως τα ChatGPT, GPT-4 και Llama έχουν δείξει τεράστια ικανότητα στην κατανόηση και παραγωγή κειμένου. Μεταξύ των πιο σημαντικών αναδυόμενων ικανοτήτων των LLMs είναι η λεγόμενη **Chain of Thought (CoT) prompting**. Η τελευταία αποτελείται από μια σειρά ενδιάμεσων βημάτων συλλογισμού που μπορούν να βελτιώσουν σημαντικά την ικανότητα των LLMs να εκτελούν σύνθετους συλλογισμούς. Επιπλέον, παρέχει ένα ερμηνεύσιμο παράθυρο στη συμπεριφορά του μοντέλου, υποδεικνύοντας τον τρόπο με τον οποίο μπορεί να έχει καταλήξει σε μια συγκεκριμένη απάντηση και παρέχοντας ευκαιρίες εντοπισμού των σημείων όπου η πορεία συλλογισμού πήγε στραβά.

Ο στόχος της προτεινόμενης εργασίας είναι να εξετάσει την εφαρμογή της μεθόδου **CoT prompting** στην ταξινόμηση προτάσεων, και πιο συγκεκριμένα στην ταξινόμηση προθέσεων (**intents**). Η εκτίμηση της πρόθεσης του χρήστη μιας υπηρεσίας είναι ένα από τα βασικά προβλήματα στα συστήματα διαλόγου προσανατολισμένα σε συγκεκριμένη εργασία. Παραδοσιακά, έχει προσεγγιστεί ως πρόβλημα ταξινόμησης που απαιτεί πολλά παραδείγματα εκπαίδευσης ανά πρόθεση και παλαιότερους ταξινομητές που εκπαιδεύονται σε ένα προκαθορισμένο σύνολο προθέσεων. Τέτοιες προσεγγίσεις δεν είναι γενικά κλιμακούμενες και δύσκολα συντηρούνται, καθώς απαιτούν ένα μοντέλο ανά πελάτη λόγω του ότι το σύνολο των προθέσεων εξαρτάται από τον πελάτη. Μια πρόσθετη αδυναμία αυτών των προσεγγίσεων είναι ότι δεν αποκαλύπτουν τον τρόπο με τον οποίο το μοντέλο ταξινομεί τις φράσεις σε κλάσεις προθέσεων. Μπορούν να εφαρμοστούν μέθοδοι επεξήγησης όπως η **LIME** ή η **SHAP**, ωστόσο δεν αποδίδουν πάντα διαισθητικές εξηγήσεις, καθώς απλώς εκτιμούν τη συμβολή κάθε λέξης της φράσης στην απόφαση του ταξινομητή, κάτι για το οποίο η μέθοδος **Chain of Thought prompting** μπορεί να δώσει λύση.

Τα LLM έχουν επιδείξει ισχυρές επιδόσεις στην ταξινόμηση κειμένου, τόσο σε πλαίσιο **0-shot** (δηλαδή χωρίς κανένα παράδειγμα εκπαίδευσης στο **prompt**) όσο και σε πλαίσιο **few-shot** (δηλαδή με ένα συγκεκριμένο αριθμό παραδειγμάτων εκπαίδευσης στο **prompt**), αλλά όχι στο ίδιο επίπεδο με παλαιότερα μοντέλα που έχουν σχεδιαστεί ειδικά για ταξινόμηση. Για να αντιμετωπιστεί αυτό το κενό στην επίδοση, προστίθεται στο **prompt** ένα σύνολο 5 υποψήφιων προθέσεων. Αυτές οι προθέσεις εξάγονται χρησιμοποιώντας έναν ταξινομητή με βάση το **BERT**. Για να μειωθεί περαιτέρω το χάσμα στην επίδοση, για κάθε μία από τις 5 υποψήφιες προθέσεις περιλαμβάνεται επίσης στο **prompt** μια περιγραφή για την πρόθεση αυτή, η οποία δημιουργήθηκε με τη βοήθεια του ChatGPT.

Όσον αφορά το μέρος της αλυσίδας σκέψης αυτής της διατριβής, δοκιμάστηκε αρχικά μία 0-shot μέθοδος αξιοποιώντας τρεις φράσεις που παράγουν Chain of Thought, «*Let's Think Step by Step*», «*Let's Take a Deep Breath and work on this Step by Step*» και μια νέα φράση, «*Show your Thoughts*». Θα δείξουμε ότι και στα δύο σύνολα δεδομένων που χρησιμοποιούνται, δηλαδή το CLINC-150 και το BANKING77, η μέθοδος που περιγράφηκε παρουσιάζει ελπιδοφόρα αποτελέσματα.

Acknowledgements

I want to thank my family who supported me my whole life on every step I ever took and helped me in every way imaginable. I also want to thank all my friends that stood by my side through the course of this degree and supported me in this ordeal. I especially want to thank my friend Renia for all the times she helped me, encouraged me and supported me. I also want to thank Panos, Deppie, and Giota, without whom this past year and a half would have been a lot worse than it actually was.

I want to thank my academic supervisor, Professor Ion Androutsopoulos both for teaching me so much during the Text Analytics course of this degree, but also for his immense support and guidance during this thesis. I also want to thank my Omilia Supervisors, Dr. Themis Stafylakis and Mr. Panagiotis Tassias for all their suggestion and directions which helped shape this thesis and bring it to life. I am also extremely grateful to my Omilia Supervisors as they provided me with a Google Colab Pro+ subscription which helped me immensely by saving me a great deal of time with my experiments.

Contents

Abstract	v
Acknowledgements	vi
1 Introduction	1
2 Related Work	5
3 Datasets and LLM	8
3.1 CLINC-150	8
3.2 Banking77	9
3.3 Llama2	11
4 Prompting Pipeline	13
4.1 Introduction	13
4.2 Top-5 Intents Extraction	14
4.3 Description Extraction	14
4.4 Technical Setup	15
4.5 Prompting Techniques	16
4.5.1 Let’s Think	17
4.5.2 Deep Breath	18
4.5.3 Show your Thoughts	20
5 Initial Results	22
5.1 CLINC-150	23
5.1.1 Chain of Thought Analysis	25
5.2 BANKING77	27
5.2.1 Chain of Thought Analysis	28
6 Few Shot Experimentation	31
6.1 Introduction	31

6.2	CLINC-150	33
6.2.1	Chain of Thought Analysis	36
6.3	BANKING77	38
6.3.1	Chain of Thought Analysis	42
7	Ablation Study & Error Analysis	45
7.1	Ablation Study	45
7.1.1	Without Chain of Thought	46
7.1.2	Without Descriptions	48
7.1.3	Without both	54
7.2	Error Analysis	56
7.2.1	CLINC-150	57
7.2.2	BANKING77	58
8	Conclusions & Future Work	61
8.1	Conclusions	61
8.2	Future Work	62
	Bibliography	64
A	Prompt Examples	68
B	Correct Responses - CLINC-150	90
C	Correct Responses - BANKING77	100
D	Error Analysis Contents - CLINC-150	109
E	Error Analysis Contents - BANKING77	115
	List of Figures	121
	List of Tables	122

Introduction

In recent years the advancements in the field of Deep Learning and especially in the Natural Language Process domain have completely changed our perception of Neural Networks. Language models have evolved from the previous generation Machine Learning models accompanied by a beam search for the next word prediction, to transformer-based models (Vaswani et al. (2017)), like Bidirectional Encoder Representations from Transformers (Devlin et al. (2019)) which revolutionized the field of NLP, to the more advanced and complicated Large Language Models like RoBERTa (Liu et al. (2019)), T5 (Raffel et al. (2020)), and GPT-like models, like GPT2 (Radford et al. (2019)) and Instruct-GPT (Ouyang et al. (2022)).

While the scientific community knew a great deal about models like these, the rest of the people did not. All these changed in December of 2022, when OpenAI released to the public for free ChatGPT (OpenAI (2022)), a Large Language Model specifically trained to behave like a chatbot. It should be noted that ChatGPT became one of the fastest-adopting platforms of all time, achieving a million users in the record time of five days. For context, Instagram needed 2.5 months to achieve 1 million downloads, whereas Netflix reached the 1 million user achievement 3.5 years after its release.

As the advancements in the field continued to rise with the creation of larger and more complex models, scientists tried to find better ways to prompt these models. One such way was few-shot prompting, where the user provides some examples of already answered questions of the same type to control to some level the way the model answers. Another such way is the reason this thesis exists and it's called Chain of Thought prompting. In its essence, in Chain of Thought prompting the user first presents some few shot examples in the Large Language Model, but with a twist: in the already presented answers there exist some intermediate reasoning steps, and not just the answer to the question. This is done in order to have some form of explainability in the model's response while increasing the performance of the model. In Fig. 1.1 we can see the same question answered both with normal prompting and Chain of Thought prompting, where the normal prompting method could not accurately calculate the answer, while the Chain of Thought example not only answered correctly, it also provided a reasoning for its answer.

This technique could prove particularly helpful in a very specific task, the task of intent classification of utterances that come from task-oriented chatbots, with the help of Large Language Models. In a scenario where it is of the utmost importance for a system to understand

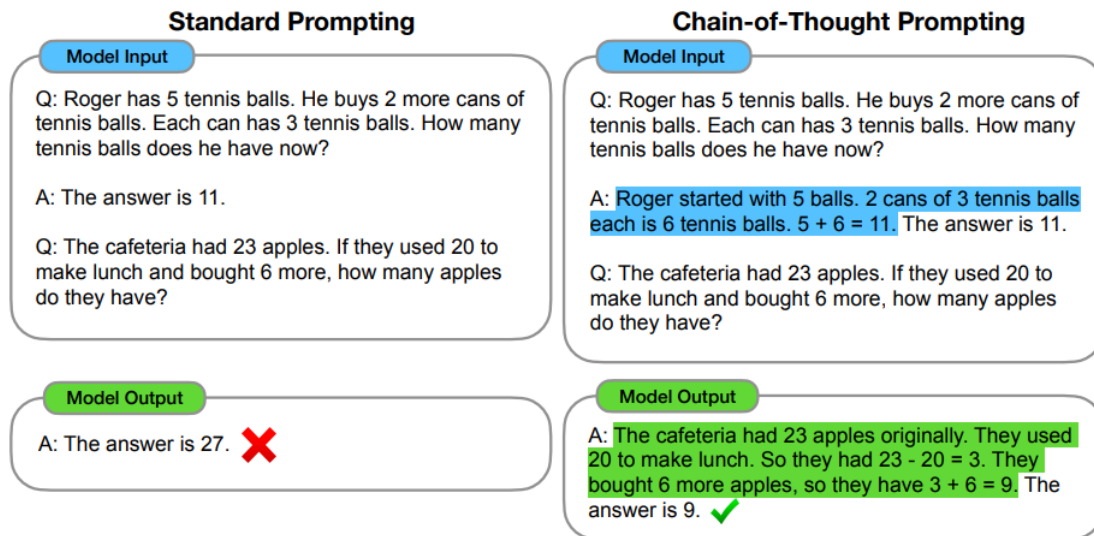


Fig. 1.1: A few-shot prompt example with Chain of Thought prompting compared with normal Prompting. Source: Wei et al. (2022)

the needs of the user that interacts with it so that it can serve the user better, the classification of the user's intent is the only way to go. Up until this point this was usually done with LSTM or simple transformer-based models, needing a very large amount of train utterances per intent while not explaining the classification. Despite the concerns raised by Turpin et al. (2023) who supported the Chains of Thought produced by the model can be extremely different when changing a few things in a prompt (e.g. reordering the possible answers in a multiple choice few-shot prompt so that the correct answer is always the same) our goal is still the same. By utilizing Chain of Thought prompting we aim to create a universal intent classification system that can be offered with very small differences to different customers while at the same time offering adequate reasoning for the predicted intent.

Our proposed approach for this thesis revolves around building a few-shot prompting pipeline which includes the 5 most probable intents as per a traditional BERT-based classifier, and their descriptions, as they were extracted using ChatGPT. The in-shot exemplars are also extracted after prompting the same model utilized for the classification of the texts. As we will prove in this thesis, this approach achieves superior performance relative to simple prompting while being relatively close to a fine-tuned version of a BERT classifier.

Besides the prompting pipeline, which we will describe in the following chapters, our main contribution with this thesis is another phrase that can be used to extract Chains of Thought from Large Language Models, *Show your Thoughts*. This phrase achieves similar or even

superior performance when compared with other such phrases like *Let's Think Step-by-Step*, both in terms of accuracy and in terms of Chain of Thought quality.

This thesis is structured as follows:

Chapter 2: Related Work

In this chapter some of the papers that inspired the work for this thesis and some more papers that are very closely related to the work of this thesis even if they are not actively utilized are presented.

Chapter 3: Datasets and LLM

This Chapter has a brief overview of the two datasets used in this thesis, CKINC-150 and BANKING77, as well as the process of creating the subsets of these datasets used in this thesis. There is also a brief description of the model that is used in the thesis, the Llama2 model.

Chapter 4: Prompting Pipeline

This Chapter gives, in great detail, an overview of the pipeline that is used for this classification task with Chain-of-Thought prompting, along with the models that are used in each stage of the pipeline, the prompting techniques that are used, and an example prompt for each one.

Chapter 5: Initial Results

This Chapter provides insights as to the results of the the experiments with the pipeline that was described in the previous Chapter, for both Chapters. For each prompting technique for both datasets, there is also a brief analysis of the Chain of Thought of some correctly classified text utterances to determine the degree to which our method achieves what we are hoping for.

Chapter 6: Few Shot Experimentation

In this Chapter, the few-shot pipeline, which extends the 0-shot pipeline, that was used in this thesis will be explored, along with the results of this technique on both datasets and the same Chain of Thought analysis that took place in the previous Chapter.

Chapter 7: Ablation Study & Error Analysis

This 7th Chapter presents a detailed Ablation Study where the parts of the prompting pipeline that are tested, like the Chain of Thought inciting phrases or the intent descriptions, are gradually removed to quantify their importance. A detailed error analysis is also included where the errors of the best methods for both datasets are reviewed and analyzed.

Chapter 8: Conclusions & Future Work

In the final Chapter of this thesis, a brief review of the conclusions of this thesis is presented along with some key ideas that could be explored in the future.

Related Work

This work is inspired by many different research areas. In this chapter, we will describe in some detail the papers that influenced us the most and we consider to be the most relevant to this project.

Wei et al. (2022) started it all with their paper "*Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*". They explore the idea, as the title puts it, of Chain of Thought Prompting which they define as a series of intermediate reasoning steps. By introducing just a few exemplars that feature these intermediate reasoning steps in the prompt for In Context Learning, models with more than 100 billion parameters seem to have massive performance gains in problems that require complex multi-step reasoning, like Arithmetic Reasoning problems, Commonsense reasoning problems, etc. Models that have less than 100 billion parameters tend to not have any performance gains by Chain of Thought Prompting, they could even lose some performance when utilizing it, as these models produce fluent yet illogical Chains of Thought. They also note that the use of hand-crafted exemplars yields a variance in performance (in some cases as high as 20%) as a result of the different annotations by different people.

Kojima et al. (2022) were inspired by Wei et al. (2022) and they tried to mitigate the large drawback that Chain of Thought Prompting presented: the need for hand-crafted exemplars that yielded a variance in the performance of the model. The way they did this was by introducing a key phrase in the prompt to incite the model to reason. This phrase is "*Let's Think Step-by-Step*". They noticed that by simply adding this phrase at the end of the prompt, the model produces logical intermediate reasoning steps. When they took these steps and inserted them at the end of the original prompt the performance of the model improved by a significant margin in all of the tasks that the original Chain of Thought Prompting was tested on, i.e. Arithmetic Reasoning, Commonsense Reasoning etc. The only drawback of their method is that the performance gain is not as great as the one with the Chain of Thought Prompting as Wei et al. (2022) suggested it, and the fact that still smaller models with less than 100 billion parameters do not have any performance gains.

Wang et al. (2023) propose a new way to boost performance using Chain of Thought Prompting, called Self-Consistency (SC). SC is what they call a "*self-ensembling*" method in the way that the model samples k different reasoning paths (i.e. k different Chains of Thought) and the most frequent one is taken as the answer (hence the self-ensembling). The greatest advantage

of this particular method is that it can be used with both hand-crafted exemplars the way Wei et al. (2022) proposed, and with zero-shot methods, like Kojima et al. (2022) proposed. In both cases, a great performance gain relative to the previous standalone methods is observed. Aggarwal and Yang (2023) did the same thing with their paper only with a slight difference as they used a probabilistic stopping criterion to find the optimal k , thus minimizing the computational cost relative to SC while retaining the performance gains that it provides.

Wang et al. (2023) propose an alternative to *"Let's Think Step by Step"*, as proposed by Kojima et al. (2022), with their method *"Plan and Solve"* (PS), where they simply replaced the key-phrase *"Let's Think Step by Step"* with their variation of it: *"Let's first understand the problem and devise a plan to solve the problem. Then let's carry out the plan and solve the problem step by step."*. Through their tests, they came to the result that their method yielded significantly better results than the method Kojima et al. (2022) proposed and they were comparable with the performance that Wei et al. (2022) reported. But, as with the previous methods, models with less than 100 billion parameters may not have any performance gains with this method.

Yang et al. (2023) present their new approach to leveraging Large Language Models as optimizers, but not in the traditional sense (i.e. like Adam Optimizer), where a model takes a prompt as input and optimizes it to yield better responses. They build upon Kojima et al. (2022) with this system, which they called OPRO. OPRO features two different Large Language models, an Optimizer model, and a Scorer model. The Optimizer model first takes as input a meta-prompt and yields 8 different candidate instructions. These instructions are then fed to the Scorer model where we evaluate the accuracy of the training set of interest (which is dependent on the task) 8 times and get 8 different accuracy scores. These scores are then added to the meta prompt, along with the candidate instructions and the process repeats itself. A by-product of this process is one key-phrase: *Let's Take a Deep Breath and work on this step by step*, which they found to be working better than *Let's Think step by step* that Kojima et al. (2022) created.

Parikh et al. (2023) present a comprehensive study on advancing the field of intent classification in conversational Natural Language Understanding systems, particularly addressing the cold-start problem for new customers. Recognizing the constraints of storage space and the need for scalability, the paper explores four innovative approaches: domain adaptation, data augmentation, zero-shot intent classification using Large Language Models, and parameter-efficient fine-tuning of instruction-finetuned language models. Through extensive experiments, the paper demonstrates that these methods, particularly parameter-efficient fine-tuning using T-few recipe on Flan-T5, significantly enhance performance in low-resource settings, even with minimal data per intent. This work stands as a substantial contribution to the field, offering

practical solutions for deploying conversational agents more effectively and efficiently across a wide range of customers and scenarios.

Magister et al. (2023) address the significant challenge of transferring the advanced reasoning capabilities of Large Language Models to smaller, more efficient models. The authors explore the effectiveness of Chain of Thought Prompting and knowledge distillation as a means to endow smaller models with the ability to perform complex reasoning tasks. Their method involves fine-tuning a smaller "*student*" model on the Chain of Thought Prompting outputs generated by a larger "*teacher*" model. The study's results are promising, indicating that this approach can significantly improve the smaller models' performance across various reasoning datasets. This research is particularly notable for its potential to make the advanced reasoning capabilities of large models more accessible and efficient, contributing to the broader field of NLP by enabling more compact models to perform complex tasks traditionally reserved for much larger models.

Loukas et al. (2023) address the challenges of using LLMs for an intent classification task on a dataset that we also used in this thesis, the BANKING77 dataset which consists of 77 fine-grained intent classes. For their research, they compared fine-tuned Masked Language Models (MLMs) with few-shot MLMs and few-shot Large Language Models LLMs. They concluded that LLMs are effective classifiers for this intent classification task, but extremely costly. For this reason, they suggested a Retrieval Augmented Generation (RAG) approach for this task which helps reduce the cost for this task by multiple times without sacrificing a lot of performance.

Datasets and LLM

3.1 CLINC-150

CLINC-150 is a dataset created by Larson et al. (2019) which addresses the issue that task-oriented dialog systems need to recognize when a query is outside their range of supported intents. Up until now, datasets created for text classification typically only define label sets that cover every example. The novelty of this dataset is the addition of a different class, called "out-of-scope" which holds queries that do not fall under any of the other categories. This is crucial because models cannot assume every query at inference time belongs to a supported intent class.

The dataset contains 150 different intents spanning 10 different domains with the 151st intent being the *out-of-scope* (*oos*) class. There exist 3 main dataset variations, one being balanced, one imbalanced where the *oos* class has many more texts than the rest of the classes, and one smaller variation, where each class has a lot fewer representatives than the full variation.

Each variation of the dataset is also split in three different subsets, a *train* subset, a *validation* subset, and a *test* subset, each being used as its name suggests.

The following table presents some statistical data for each of CLINC-150's balanced set data splits, i.e. the train split, the validation split, and the test split.

Split	Length	Level	Mean	std	min	max	median
Train	15,100	Character	39.89	15.25	2	136	38
		Word	8.33	3.19	1	28	8
Validation	3,100	Character	39.71	16.5	2	114	37
		Word	8.28	3.42	1	24	8
Test	5,500	Character	40.17	15.61	2	125	38
		Word	8.29	3.19	1	25	8

Tab. 3.1: Statistics in both Character Level and Word Level for CLINC-150. The mean lengths are shown along with their standard deviation and the minimum, maximum, and median lengths.

In the following table, we will also present some example text utterances along with their assigned intent class.

text	intent
i need anniversary dinner added to my schedule for march 7	calendar_update
is there traffic to work	traffic
hey, slow down and take your time	change_speed
my card's magnetic strip is not working and i want to notify them	damaged_card
what are you able to do	what_can_i_ask_you
please take final exams off my calendar for may 3	calendar_update
i recently took a new job and need help with my 401k rollover	rollover_401k
the status of my day off request, please	pto_request_status
how long until i need to change my oil	oil_change_when
what is your version on the meaning of life	meaning_of_life

Tab. 3.2: Examples of text utterances from CLINC-150 and their assigned intent class.

In this thesis, we have altered the dataset in a way to better suit our needs and limitations. First and foremost we have completely dropped the *oos* class from all of the data splits as it is not in the scope of this project to detect out-of-scope queries. For all our experiments (which we will describe in great detail in the following chapters) we use a subset of the *test* split which we have created by randomly selecting 5 texts from each intent which brings down the length of the test dataset from 5500 texts to just 750. This was done because the inference times of the model we utilized were simply too large for 5500 texts.

3.2 Banking77

Banking77 is a dataset created by Casanueva et al. (2020) and differs massively from CLINC-150 in the sense that it features utterances only from the banking domain which allows for a deeper understanding of the specific types of queries and needs within the industry. Its main aim is to address the complexities and nuances of intent classification in customer-based interactions, like task-oriented dialog systems.

It features 77 unique fine-grained intents with 13,083 utterances that come from various customer service queries. The fine-grained nature of the intents makes the dataset more realistic and challenging as it moves away from oversimplified intent detection tasks to better reproduce the variety and complexity of actual customer interactions in a commercial setting.

Unlike CLINC-150 which had different variations based on the distribution of instances across the classes, BANKING77 has only one variation and it is an unbalanced one since some intent classes have a lot more instances than others. The presented violin plot shows this imbalance across the classes. This plot was created to have the number of instances of each intent class (therefore the dataset had 77 observations). As we can see, most of the counts are somewhere in the middle of the y-axis or around 120 texts per intent class.

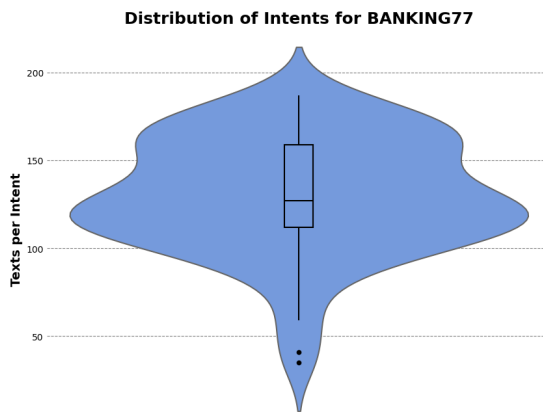


Fig. 3.1: Violin plot of class counts for BANKING77

Let us now present some example texts for this dataset too.

text	intent
What documentation is needed for the identity check?	verify_my_identity
What is your maximum top up limit per day?	top_up_limits
Why isn't my cheque deposit showing on my account? I thought it would be faster and would like to look into it and help.	balance_not_updated_after_cheque_or_cash_deposit
I am entering my passcode but getting an error.	passcode_forgotten
Why is my cash request and the amount I received different?	wrong_amount_of_cash_received
What do I do if the machine rejects my card?	card_not_working
Explain pending transactions.	pending_cash_withdrawal
is there a limit on top ups	top_up_limits
I wish to be able to top up with cash.	top_up_by_cash_or_cheque
Why do I see an extra €1 charge on my statement?	extra_charge_on_statement

Tab. 3.3: Examples of text utterances from BANKING77 and their assigned intent class.

Split	Length	Level	Mean	std	min	max	median
Train	10,003	Character	59.47	40.86	13	433	47
		Word	11.95	7.89	2	79	10
Test	3080	Character	54.23	34.65	13	368	45
		Word	10.95	6.69	2	69	9

Tab. 3.4: Statistics in both Character Level and Word Level for BANKING77. The mean lengths are shown along with their standard deviation and the minimum, maximum, and median lengths.

As we did with CLINC-150, here too, to save time in inference we kept a balanced subset of the test dataset featuring 5 utterances per intent, bringing the total length of the test dataset to 385 texts.

3.3 Llama2

Llama 2 (Touvron et al., 2023) is a family of Large Language Models created and released as open-source models by MetaAI, the research division of the Meta company. The models come in 4 sizes (with 7, 13, 34, or 70 billion parameters) and 2 variations (either stock or fine-tuned for chatting purposes). The reality is that even MetaAI wanted to release the 34B model in both variants, at the time of writing this thesis this has not happened yet.

Each model of the Llama 2 family of models was trained on the same training data all of which come from publicly available sources and do not include any amount of data from Meta's products or services, like Facebook, Instagram, etc. The authors do not specifically say what this dataset includes but the authors mentioned that they took extra caution to remove data from sites that feature personal information, which brought the total size of the training dataset to 2 trillion tokens. The authors hope that with a training dataset of this size, the knowledge of the models will expand when compared with the knowledge of the previous generation of Llama models, and the hallucinations will be reduced to a minimum.

The architecture of every model of the Llama family is identical to one another with only the model size changing between different models. It remains mostly the same as in the previous generation Llama model. It is built upon the standard Transformer architecture (Vaswani et al., 2017). They also used a pre-normalization technique called RMSNorm (or Root Mean Squared Normalization) that normalizes the input of each transformer sub-layer. This way the 2 trillion training tokens and the multiple billions of parameters become a little easier to handle in terms of memory and computational time. They also used the SwiGLU activation function and Rotary Positional Embeddings (Su et al., 2021) which will help the models understand the importance of the position of each word in a different way than ordinary positional embeddings use as it encodes absolute positional information with a rotation matrix and naturally incorporates explicit relative position dependency in self-attention formulation. The primary differences between the first generation Llama model and this one are the fact that the first generation was trained on fewer data (1 trillion for the 7B and 13B models and 1.4 trillion for the 33B and 70B models), the longer context length of Llama 2 (4k tokens vs 2k tokens for the original

Llama) and the use of a Grouped Query Attention (Ainslie et al., 2023) which will help with the inference scalability of the model.

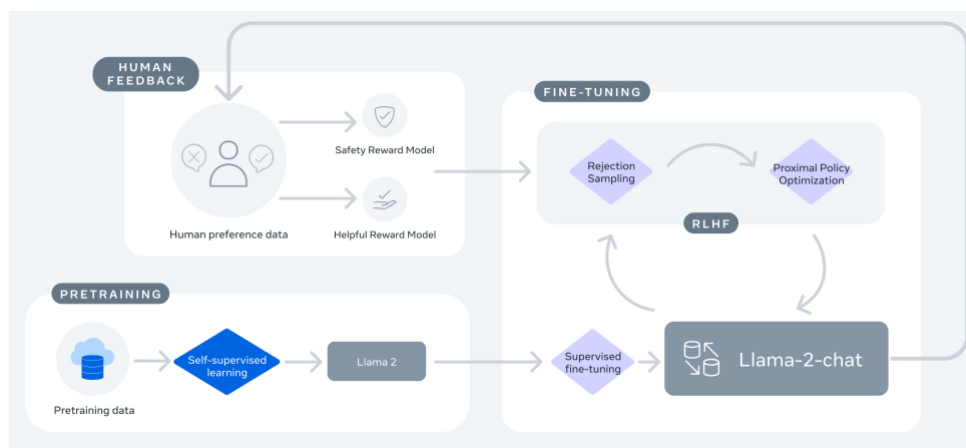


Fig. 3.2: The fine-tuning process MetaAI used on Llama. Source: Touvron et al. (2023)

One of the areas that the authors emphasized the most was the fine-tuning process they utilized to create the chat variants of the Llama model, which can be seen on Fig. 3.2. For this fine-tuning process, they used multiple techniques including Supervised Fine Tuning (SFT) where they used hand-crafted data from different sources to perform some safety control and forbid Llama from answering in a harmful way. They also used a process called Reinforcement Learning with Human Feedback (Ouyang et al., 2022) where an annotator checks the responses of the Llama model and a reward system either rewards or punishes the model. Finally, they also used a novel method called Ghost Attention which allows the model to follow instructions over multiple turns of dialog without the need to repeat the instruction. This in turn helps with the dialog flow over multiple turns.

After testing the model both on popular benchmarks and against other chat models like ChatGPT they found that all models performed admirably for the category they were in. The 7B model outperformed every other 7B model the authors tested in all but one test, the 13B outperformed the previous generation 13B variant on every single benchmark and in many cases outperformed larger models like the 33B Llama 1, MPT 30B and Falcon 40B. The larger model sizes performed admirably on these same benchmarks, which was expected after the results we previously discussed. Especially the 70B variant performed very similarly with much larger LLMs on the same benchmarks, like GPT3.5, GPT4, PaLM and PaLM-2-L, which is extremely important since all other models' parameters range in the hundreds of billions.

Prompting Pipeline

4.1 Introduction

In this chapter, we will present the pipeline we used to perform the classification task using Chain of Thought prompting. Since Large Language Models are not the traditional models used for text classification, and most of the time they are nowhere near the state-of-the-art performance, we have incorporated to our pipeline elements that will make it easier for our model to complete the task it was assigned. Below is a figure that presents visually the pipeline we created for the purposes of this thesis.

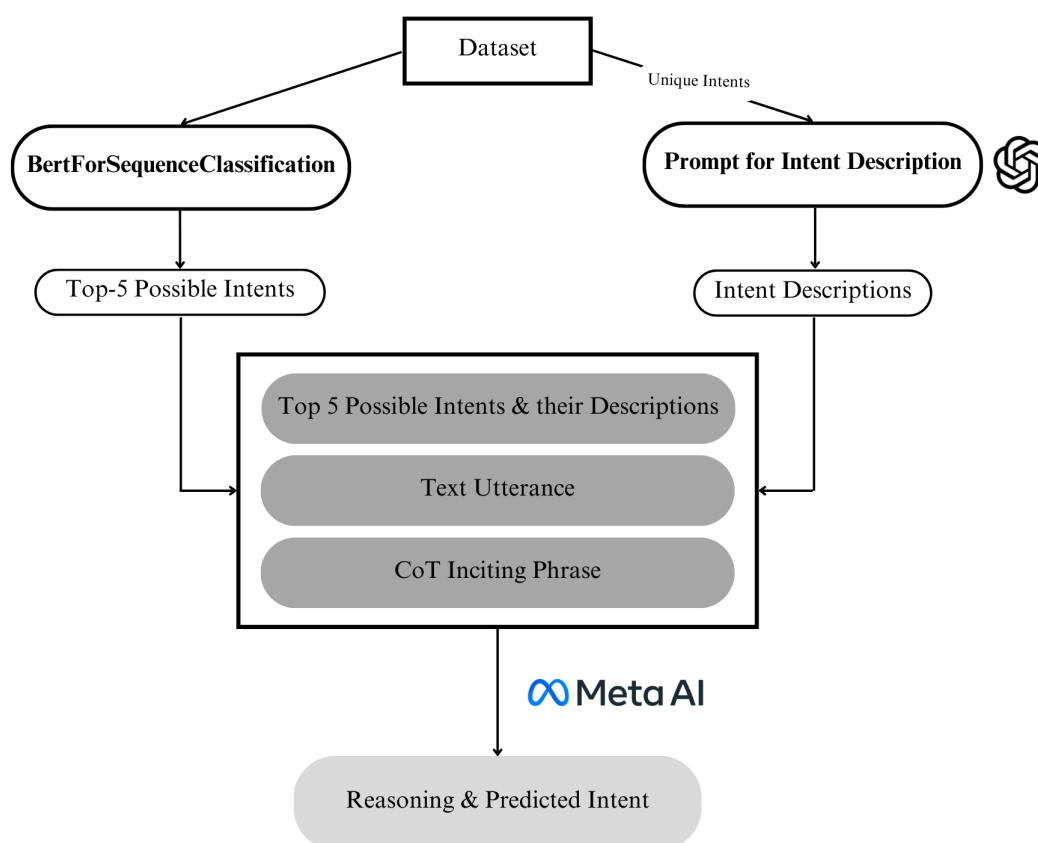


Fig. 4.1: An overview of the pipeline used in this thesis

4.2 Top-5 Intents Extraction

The first phase of our methodology involves the deployment of a classifier that represents the prior state-of-the-art, utilizing the training subsets of the dataset. For this critical task, we selected the BERT classifier, available to us through the *Transformers*¹ Library from Huggingface, and more specifically the *TFBertModelForSequenceClassification*² function. The architecture of this model consists of a BERT-base model with 12 Transformer blocks, with each block consisting of 12 attention heads. The 12 transformer blocks are followed by a dropout layer with a rate of 0.1 to mitigate overfitting and a classification layer with a softmax activation function, the nodes of which are equal to the number of classes within each dataset.

To assess the performance of this classifier, we used two metrics: accuracy and top-5 accuracy. The latter reveals the frequency with which the correct intent class ranks among the five most probable predictions as determined by the softmax activation function.

For the loss function, we chose the Sparse Categorical Cross Entropy loss, paired with the Adam Optimizer, a learning rate of $2e^{-5}$, and an epsilon value of $2e^{-8}$. We trained this classifier for 15 epochs on a single TPUv3.8 provided by Google Colab, during which we saved the best weights whenever a superior top-5 accuracy score was achieved.

We use this BERT classifier for two reasons. Firstly, it serves as a contemporary benchmark against which we can measure our results, providing a clear indication of our successes and areas for refinement. Secondly, it enables us to extract the top 5 predictions for the test dataset, a crucial step towards the creation of the prompts used in our final system. We will elaborate further on the use of these top 5 predictions in subsequent chapters.

4.3 Description Extraction

The next step of our pipeline is the extraction of the intent descriptions. For this, we employed a straightforward yet somewhat labor-intensive method that drew inspiration from the scholarly work of Bouzaki (2023).

¹<https://huggingface.co/docs/transformers/en/index>

²https://huggingface.co/docs/transformers/model_doc/bert

Our method involved interacting with the web interface of ChatGPT, specifically the GPT-3.5 version, utilizing a prompt crafted by Bouzaki (2023). This prompt includes two hand-crafted intent class descriptions designed to guide ChatGPT's responses, subsequently prompting it to generate a comparable sentence describing the intent class of our choice. The original prompt is as follows:

The intent "translate" represents ways to ask for the translation of a word or phrase in a language different from the one being used. The intent "find_phone" refers to a request for assistance in locating a lost cell phone or a question about its current location. Create only one similar sentence with a description for the intent [x], without mentioning any example of it.

where [x] is the intent that we want to extract its description each time.

Given that both our research and Bouzaki (2023) are grounded in the CLINC-150 dataset, we saw no necessity to alter this prompt, as it has proven to be quite effective. However, considering the BANKING77 dataset, we devised a second prompt, closely modeled after the first, to cater to this dataset's unique intents. This second prompt also includes two tailor-made intent class descriptions to steer ChatGPT's response. We then ask it to provide a similar descriptive sentence for the intent class we are examining.

The exact wording of the second prompt is as follows:

The intent "card_arrival" refers to a request for assistance in tracking a card or a question about its current location or time of arrival. The intent "card_payment_wrong_exchange_rate" represents ways to express frustration with the wrong exchange rate between different currencies being applied on a internet purchase. Create only one similar sentence with a description for the intent [x], without mentioning any example of it.

where [x] is the intent that we want to extract its description each time.

In the next sections, we'll elaborate on the precise manner in which we employ these extracted descriptions.

4.4 Technical Setup

Our chosen methodology for crafting prompts might appear somewhat unconventional at first glance, yet it's fundamentally the most effective strategy. Recognizing that Large Language

Models lack genuine cognitive abilities and are prone to generating misleading or hallucinatory content, it becomes necessary to provide them with extra information. To this end, we incorporate the top five intents as predicted by BERT at the beginning of each prompt. These top five intents are accompanied by their description as they were crafted by ChatGPT. This was done to reduce the number of possible intent classes that the model has to choose from as both datasets feature an enormous amount of intents (150 and 77 respectively), thus reducing the size of each prompt and effectively reducing inference times.

The model we chose to do this was the 13B version of the Llama2³ model, created by MetaAI. This model was chosen among others of similar size because it surpassed similar models in terms of performance (at least at the time when this thesis began) in almost every task. As the size of this model is extreme and cannot fit in any available GPU, we loaded the model using a 4-bit quantization method (Jacob et al., 2017) which reduces the model size enough to fit in a T4 GPU (available on the free tier of Google Colab), without reducing its capabilities and *handicapping* the model.

The inference was done on a A100 GPU, available to us by the Pro+ Tier of Google Colab, utilizing the *text generation*⁴ pipeline by *transformers* with this parameter specification: temperature = 0.7, top_p = 0.95, max_new_tokens = 200 and repetition_penalty = 1.1. On this specific pipeline, we fed each test prompt we are about to describe with a batch size of 1, as having a batch size greater than 1 could result in memory overflow thus terminating the inference procedure.

4.5 Prompting Techniques

Our prompting strategies that are designed to elicit reasoning from the Llama model are three. For the purposes of this thesis, we will name them *Let's Think*, *Deep Breath* and *Show your Thoughts*.

³<https://llama.meta.com/>

⁴https://huggingface.co/docs/transformers/en/main_classes/pipelines

4.5.1 Let's Think

This innovative prompting strategy is the work of Kojima et al. (2022) who drew inspiration from the original Chain of Thought Prompting paper (Wei et al., 2022). Their objective was to emulate the outcomes of the original study but through an automated process that eliminates the necessity for hand-crafted few-shot exemplars. The methodology they devised can be seen in **Fig. 4.2**. They begin by appending the phrase *Let's Think Step by Step* to the end of the initial prompt, a deliberate addition designed to encourage the Language Model to articulate its reasoning process. Following this, the response generated by the model is integrated back into the original prompt as an intermediate step, and this augmented prompt is then re-submitted to the model to elicit an answer.

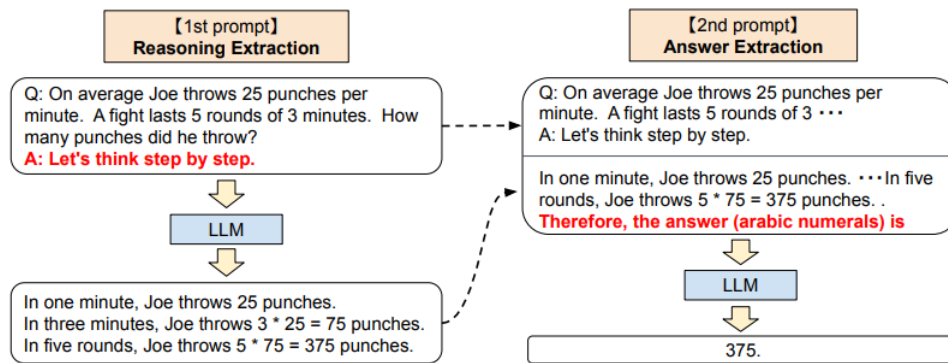


Fig. 4.2: The two-stage prompting method of 0-shot Chain of Thought prompting. Source: Kojima et al. (2022)

Even though this prompting technique was proven useful for Kojima et al. (2022), it presents a practical challenge in our case, primarily due to the relatively slow inference speed of Llama2 and our lack of resources. Undergoing two iterations through the model to arrive at the desired answer is a luxury we cannot afford. Consequently, we adapted the method proposed by Kojima et al. (2022) to extract both the Chain of Thought and the final answer within a single prompt. The way we did this is by asking the model to figure out the intent of the user when they wrote the input text while adding the *Let's Think Step by Step* phrase after the question. With this process, we eliminate the necessity for a second prompt, effectively halving the inference time compared to Kojima et al. (2022) methodology, thereby enhancing the efficiency of our process.

Below we present a full prompt for each dataset to show exactly what we meant in the previous paragraphs.

We have the following set of intents along with their descriptions:

- * **no**: The intent "no" signifies a straightforward refusal or a negative response to a question or statement, indicating a lack of agreement or affirmation.
- * **what_can_i_ask_you**: The intent "what_can_i_ask_you" pertains to seeking guidance or suggestions on the types of questions or topics that can be posed.
- * **what_are_your_hobbies**: The intent "what_are_your_hobbies" involves prompting someone to share information about their personal interests and activities, offering insights into the individual's leisure pursuits and pastimes.
- * **insurance**: The intent "insurance" pertains to inquiries or requests related to obtaining information about insurance coverage or the specifics of the insurance plan of the user.
- * **calculator**: The intent "calculator" involves users seeking access to or information about a digital or physical calculator, expressing a need to perform mathematical calculations or inquiring about the functionalities of a calculator tool.

A user wrote the following text: "can i know what type of things you can help me with". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's think step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

As we can see, the top 5 intents, as they were extracted using BERT, were inserted at the beginning of the prompt, along with their respective descriptions, as they were created using ChatGPT, for our model to choose one of them as the correct one.

After these possible intents of the user's utterance, we present the utterance itself on the model and we ask it to return to us the intent of the user when they wrote this text.

Finally, we use the *Let's Think Step by Step* phrase to elicit reasoning from the model. We accompany this phrase with some behavior-setting phrases that we found to work, asking the model to answer in a single sentence, do not speculate about the answer (i.e. not use words like *likely*, *possibly* etc) and include in its answer the intent name exactly as is written above.

4.5.2 Deep Breath

Let's take a Deep Breath and work on this Step by Step was initially conceptualized by Yang et al. (2023). This prompt emerged as an enhanced version of the *Let's Think Step by Step* strategy,

originally introduced by Kojima et al. (2022). The empirical investigations of the Google Deepmind engineers who worked on this prompting method revealed that this particular prompting strategy outperformed its counterparts, yielding superior results within their Prompt Optimization framework. Encouraged by these findings, we decided to incorporate this strategy into our prompting techniques' choices.

We used this prompting strategy in the same way we did with the *Let's Think* one. Specifically, we appended the *Let's take a Deep Breath and work on this Step by Step* phrase at the end of our prompt, along with the same behavior-setting phrases as before. This way we incite the model to perform the necessary intermediate reasoning steps and at the same time perform the classification of the utterance to one of the 5 possible intent classes.

We will now present a full prompt to fully visualize the previous description of our prompt.

We have the following set of intents along with their descriptions:

- ***pending_transfer**: The intent "pending_transfer" refers to inquiries or concerns about the status of a financial transaction that has been initiated but not yet completed, often seeking clarification on the expected completion time or reasons for the delay.
- * **reverted_card_payment?**: The intent "reverted_card_payment?" pertains to inquiries or concerns regarding transactions that have been canceled or reversed on a card, often seeking clarification on the reasons, processes, or implications of such reversals.
- * **transaction_charged_twice**: The intent "transaction_charged_twice" corresponds to inquiries or concerns regarding a situation where a user's account shows that a single transaction has been processed and debited more than once, leading to confusion and a need for clarification or resolution.
- * **request_refund**: The intent "request_refund" refers to an inquiry or action taken by a user to seek the return of funds for a transaction or service that did not meet their expectations or requirements.
- * **Refund_not_showing_up**: The intent "Refund_not_showing_up" pertains to inquiries or concerns regarding the visibility or status of a processed refund that has not yet appeared in the individual's account or financial statements.

A user wrote the following text: "How long will it take to get my refund". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

Again, the top 5 intents that were extracted using BERT are at the beginning of the prompt, along with their respective descriptions, as they were created using ChatGPT. These 5 classes are presented as the only possible classes for our model to choose from.

In the next part of the prompt, the user's utterance is presented to the model in the same manner as the *Let's Think* method. We then request the model to determine and relay back to us the specific intent of the user at the time of writing their text.

Last but not least, we insert the *Let's Take a Deep Breath and Work on this Step by Step* phrase, the most important part of our prompt, to elicit reasoning from the model. This phrase is followed by the same behavior-setting phrases as before to somewhat control the output of the model.

4.5.3 Show your Thoughts

Show your Thoughts is our novel approach to a prompt strategy that encourages Large Language Models to exhibit their reasoning processes. When reading the relative work for these prompting techniques we noticed the simplicity of these phrases, which by simply asking the model to think step by step or work on a problem step by step, can extract the reasoning behind the model's decision. We decided to take a similar path and just keep it simple. The simplest way to ask for the thoughts of someone is to ask them to show their thoughts. Thus the *Show your Thoughts* prompting strategy was born, which we pit against these two strategies for simple Chain of Thought extraction.

Similarly to the previous two strategies, we append *Show your Thoughts* at the end of our prompt, along with the same behavior-setting phrases we used before. The goal is to have the model not only provide an answer but also reveal the reasoning path it utilized to arrive at that particular answer.

We have the following set of intents along with their descriptions:

- * **schedule_maintenance**: The intent "schedule_maintenance" involves seeking help or information regarding the arrangement of upcoming maintenance activities for a car.
- * **gas_type**: The intent "gas_type" involves seeking information about the specific type or grade of fuel required for a vehicle or a related inquiry about available fuel options.
- * **oil_change_when**: The intent "oil_change_when" involves seeking information or recommen-

dations regarding the appropriate timing or intervals for performing an oil change in a vehicle, considering factors such as mileage, driving conditions, and the specific requirements of the vehicle manufacturer.

* **oil_change_how**: The intent "oil_change_how" pertains to inquiries seeking guidance or instructions on the process of performing an oil change for a vehicle, including steps and recommended tools.

* **shopping_list**: The intent "shopping_list" involves requests or actions related to creating, managing, or obtaining information about a list of items to be purchased during a shopping activity, whether it's in-store or online.

A user wrote the following text: "put together a list of instructions for me on how to change the oil in my car". The intent of this text is definitely one from the five intents in the previous set. What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

In the above example of a prompt using our *Show your Thoughts* prompting method, we can see the similarities with the prompts of the *Let's Think* and *Deep Breath* methods.

The prompt begins by incorporating the five primary intents as identified through BERT, each accompanied by detailed descriptions that we crafted using ChatGPT. These intent classes are introduced as the only possible intent classes for the text utterance that follows.

Next in the prompt, we display the user's utterance exactly as it appears in the dataset we have. Following this, we prompt the model to classify the utterance into one of the five intent categories that were previously outlined.

The final element of the prompt is the phrase *Show your Thoughts*, which will hopefully make the model provide a logical reasoning for its decision. Following *Show your Thoughts* we include the same behavior-setting phrases from the *Let's Think* and *Deep Breath* methods.

More prompts from each prompting strategy can be found in Appendix A

Initial Results

Following the discussion of the previous chapter we employed the text generation pipeline outlined earlier and applied the three prompting techniques described before to our intent classification task with Chain of Thought. Recognizing the inherent variability in Large Language Model responses, we conducted each experiment three times for each prompting technique. This approach was crucial because Large Language Models, not unlike traditional models, can yield different predicted intent classes for the same text input across separate runs. However, the distinction with Large Language Models is more pronounced given that their outputs are textual responses from which the intended class must be extracted, unlike traditional models where the primary output is the predicted intent itself. To account for this variability and gain a more comprehensive understanding of the performance of each prompting technique, we repeated each experiment three times.

The effectiveness of each prompting method is assessed with the accuracy metric. This metric is the most suitable given the structure of our test datasets, which are balanced in nature, containing an equal number of samples—five from each intent category. This balance enhances the reliability and relevance of the accuracy metric as a measure of performance. In scenarios where datasets are unbalanced, the accuracy metric would not offer the same level of reliability, and different metrics, like F1-Score would be employed.

Given that the output of this project is textual rather than merely labels or probabilities, determining accuracy for the test datasets proved to be more complex than initially anticipated. A detailed examination of a subset of the responses revealed that the model's replies often diverged from the exact intent names as presented in the prompts. Variations included the removal of underscores, spelling mistakes, the insertion of additional words within intent names, or even the generation of entirely new intent names. Consequently, relying on a rule-based approach for calculating accuracy was deemed impractical, as it became evident that there would always be an outlier response that such a set of rules would fail to accurately categorize.

In light of these challenges, we adopted a more labor-intensive yet accurate methodology. This involved manually reviewing each response generated by the model across all runs and prompting techniques to identify the predicted intent. This process allowed us to account for and adapt to every error on a case-by-case basis. We established a flexible set of rules for evaluating these errors, where if the model's prediction was essentially correct but included

minor errors—such as spelling mistakes (e.g., "recognised" instead of "recognized", at least in the context of the correct spelling for each dataset) or the use of a different form of a word (e.g., "top" instead of "topping")—we would still regard the predicted intent as accurate. Below, we provide several examples to clarify our approach and demonstrate how we managed these discrepancies.

class name in response	actual class name	correct	reason
top_up_by_card	topping_up_by_card	True	spelling error
car rental	car_rental	True	omitted _
direction	directions	True	omitted letter
pending_payment	pending_card_payment	False	word missing
top_up_by_cheque	top_up_by_cash_or_cheque	False	word missing
what_is_your_age	how_old_are_you	False	complete change
cancel_whisper_mode	whisper_mode	False	word added
correct	yes	False	complete change
what	what_song	False	word missing

Tab. 5.1: Examples of mistakes made by Llama on inference and how we classified them.

These are just a few examples to illustrate the mistakes made by the model in the class names. As we can see there are cases where the model omitted a word from the original class name to the predicted class name, it added a word to the original class name as it thought it better described the true intent of the user, or it completely changed the name of the class.

5.1 CLINC-150

We now present our findings from applying various prompting techniques to the initial dataset, CLINC-150, which we presented in Section 3.1. To provide some context, the CLINC-150 dataset is comprised of 150 distinct intent classes that span across 10 different domains. For our testing, the dataset utilized contains 5 texts for each intent class. This configuration results in a total of 750 texts within our test dataset.

Prompting Technique	1st run	2nd run	3rd run	Average	St. Dev.
Deep Breath	85.33	86.26	86.26	85.95	0.54
Let's Think	85.73	83.73	86.93	85.46	1.61
Show your Thoughts	86.00	88.13	86.40	86.84	1.13
BERT	96.00				

Tab. 5.2: Accuracy of CLINC-150 across 3 runs and their average.

We observed that each of the three techniques yielded reasonably high results, especially if one considers the 0-shot setting of this line of experiments, consistently achieving over 85% accuracy in all but one instance. However, our novel method, *Show your Thoughts*, distinguished itself by surpassing the 86% mark in all three runs, with an average accuracy of 86.84% and a standard deviation of 1.13. Specifically, the second trial recorded the highest accuracy at 88.13%, while the first and third trials reported an accuracy of 86% and 86.4%, respectively. This indicates a minimal variability of just 2.13 percentage points between the best and worst performances observed.

Deep Breath achieved an average accuracy of 85.95%, trailing by a mere 0.89 percentage points behind the leading method, *Show your Thoughts*, while also achieving the best standard deviation of the three methods, 0.54. Remarkably, *Deep Breath* demonstrated consistent performance, with both the second and third runs yielding an identical accuracy of 86.26%, while the initial run registered an accuracy of 85.33%. This method exhibited the least variability among the three prompting techniques under evaluation, with all outcomes falling within a narrow 0.93 percentage points range. This consistency suggests that *Deep Breath* may be a more stable prompting approach, offering outcomes with higher certainty regarding their accuracy, even if it does not always guarantee the best performance of the model.

Securing the third position among these techniques is *Let's Think*, which narrowly missed the second spot by a mere 0.5 percentage points, recording an average accuracy of 85.46% and a standard deviation of 1.61. This performance is also 1.38 percentage points lower than that of *Show your Thoughts*. The relatively modest score of *Let's Think* could be due to the significant variability observed across its runs. Specifically, the first run achieved an accuracy of 85.73%, the second run dropped to 83.73%—marking the lowest performance across all prompting techniques—and the third run rebounded impressively to 86.93%, which stands as the second highest performance across all nine evaluations. This resulted in a variability span of 3.2 percentage points between the best and worst runs, a figure that is relatively high compared to the variations seen with the two preceding techniques.

What remains to be seen is the comparison between these techniques and the BERT classifier we used to extract the top 5 intents per text. Recall that while our main monitored metric on that classifier was indeed the top-5 accuracy, we also monitored the plain accuracy of the classifier. BERT produced a 96% accuracy on the whole of the test dataset, which is almost 10 percentage points greater than our best result so far, making all our techniques' results kind of irrelevant relative to that great accuracy result. Even though there is no improvement when compared with the BERT classifier, this difference in performance is a very useful insight

and we cannot simply overlook it. We have to remember that while high accuracy is one of our focused areas, our main focus is the creation of logical and correct Chains of Thought from our model to have explainable results, something that is not possible with a plain BERT classifier.

5.1.1 Chain of Thought Analysis

We will now showcase some example responses generated by our Llama model, originating from utterances that were accurately classified. This is because this series of experiments will not be a part of the final system and we just want to showcase the potential of these prompting techniques. Our evaluation will focus on the quality of the Chain of Thought processes exhibited, if any, within these responses. For each prompting technique evaluated, we will present a total of 3 examples, distributing them equally with 1 example from each distinct test run. We also note that, for the inferences made in this subsection we took into account the additional examples that can be found in Appendix B.

Show your Thoughts

text: both could be

response: the intent of the user when they wrote the text "both could be" is "maybe" because the user is expressing uncertainty and exploring multiple possibilities.

text: i was at publix trying to buy some food but my card wouldn't work

response: the intent of the user when they wrote the text "i was at publix trying to buy some food but my card wouldn't work" is most likely "card_declined" because the user is seeking clarification or assistance due to their card being declined during a transaction.

text: could you use the female voice

response: the intent of the user when they wrote the text "could you use the female voice" is "change_accent", as it pertains to a request related to modifying the voice assistant's accent or language pronunciation.

From these 3 responses along with the ones presented in Appendix B, we can see that Llama produces accurate and eloquent Chains of Thought. This is of the utmost importance since it proves that our novel approach can indeed produce Chains of Thought while offering clear and comprehensible insights into why the model categorized each utterance into the specific intent class it did. Thus, we can confidently regard this technique as reliable for producing explainable and accurate classification results.

Deep Breath

text: can you tell me what hobbies the ai has

response: the intent of the user when they wrote the text "can you tell me what hobbies the ai has" is "what_are_your_hobbies", as it matches the description of the intent "seeking information about the personal interests and activities of the ai".

text: can i increase the credit limit on my mastercard

response: the intent of the user when they wrote the text "can i increase the credit limit on my mastercard" is "credit_limit_change".

text: what would be the best way to deal with lost luggage

response: the intent of the user when they wrote the text "what would be the best way to deal with lost luggage" is "lost_luggage".

In contrast to *Show your Thoughts*, *Deep Breath* does not seem to produce Chains of Thought as much, as in most of the responses Llama merely echoes the provided text and assigns a predicted intent class without further elaboration. The single result that features a Chain of Thought is eloquent thought and follows a correct logical path, as the ones created using *Show your Thoughts*. But, even though the responses do not include Chains of Thought as much, the high accuracy rate of this prompting technique could be seen as indicative of the model's ability to effectively distinguish between utterances of one intent and those of another. Thus, it appears that, at least within the context of this dataset, the method does not perform exceptionally well, particularly in terms of generating Chains of Thought.

Let's Think

text: i want to talk english

response: the intent of the user when they wrote the text "i want to talk english" is "change_language" because the user is seeking guidance on how to switch to speaking english on a device or platform.

text: could you find out if my bank charges transaction fees on my american card if use internationally

response: the intent of the user when they wrote the text "could you find out if my bank charges transaction fees on my american card if use internationally" is "international_fees".

text: find a decent rental care for next week at the philadelphia airport

response: the intent of the user when they wrote the text "find a decent rental care for next week at the philadelphia airport" is most likely "car_rental" because the text mentions renting a car and providing specific location and time frame information, which aligns with the description of the "car_rental" intent.

Here we observe a more nuanced set of outcomes. In just 1 out of the 3 responses, there is an absence of a produced Chain of Thought, with the model merely reiterating the input text alongside a predicted intent class. Conversely, the remaining 2 responses did manifest Chains of Thought that were both accurate and comprehensive, effectively illuminating the model’s rationale behind classifying each input utterance into its respective intent class. Although this approach ranked as the least effective among the three prompting techniques evaluated in this thesis, the presence of accurate and explanatory Chains of Thought, especially when one considers the additional examples of Appendix B, renders it partially reliable. This reliability pertains to both its classification accuracy and its capacity to produce Chains of Thought, offering a mixed yet insightful performance.

5.2 BANKING77

We now present the outcomes of our prompting techniques as applied to the BANKING77 dataset, which we presented in Section 3.2. This particular dataset is comprised of 77 distinct intent classes, all derived from the banking domain. To provide some context, the test subset we compiled for this dataset comprises 5 utterances for each intent class, culminating in a total of 385 text entries for our evaluation.

Prompting Technique	1st run	2nd run	3rd run	Average	St. Dev.
Deep Breath	74.54	72.20	73.50	73.31	1.17
Let’s Think	71.68	69.09	73.24	71.33	2.09
Show your Thoughts	72.98	73.76	71.42	72.72	1.19
BERT	94.00				

Tab. 5.3: Accuracy of BANKING77 across 3 runs and their average.

In this dataset, we observe a different picture than the one we observed with the CLINC-150 dataset. The standout method in this analysis is *Deep Breath*, which achieved both the highest single-run accuracy across all examined techniques with a 74.54% on its initial attempt with subsequent accuracy scores of 72.2% and 73.5% on its second and third attempts, respectively. This leads to an overall average accuracy of 73.31% and a standard deviation of 1.17.

The variability in *Deep Breath*’s performance metrics is particularly significant when compared with its behavior on the CLINC-150 dataset. The observed range of results on this dataset is approximately 2.34 percentage points, a notable increase from the sub-1 percentage point range observed in CLINC-150.

The silver medal is awarded to *Show your Thoughts*, which closely trailed behind *Deep Breath* in performance. Across three runs it achieved its highest score of 73.76% on the second attempt. The first and third attempts yielded scores of 72.98% and 71.42%, respectively. These results lead to an average accuracy rate of 72.72%, narrowly missing the top position held by *Deep Breath* by a marginal difference of 0.5 percentage points, and a standard deviation of 1.19. It's also worth noting that the consistency of this technique's performance mirrors that of *Deep Breath*, with all scores falling within a range of 2.34 percentage points.

Once again, *Let's Think* found itself at the bottom of the rankings among the three techniques, recording the lowest performance of all nine attempts with a 69.09% accuracy on its second run. In contrast, its first and third attempts showed somewhat improved outcomes, achieving 71.68% and 73.24% respectively. These performances contribute to an average accuracy rate of 71.33%, which is over 1 percentage point lower than that of *Show your Thoughts* and nearly 2 percentage points less than *Deep Breath*, while the standard deviation was 2.09. Notably, *Let's Think* exhibited significantly greater variability in its results compared to the other two techniques. This is evidenced by a spread in outcomes exceeding 4 percentage points.

The comparison with BERT offers a different perspective on the intent classification task. BERT significantly outperforms the LLM-based classifier, achieving an accuracy rate of approximately 92% across the entire test dataset and a 94% on the same test subset that is used for inference with Llama. This represents a substantially wider gap than what was observed in the CLINC-150 dataset, with a difference of more than 20 percentage points in accuracy between the BERT classifier and the Llama-based approach. Despite our primary objective being to enhance explainability in predicted intents through the Chain of Thought methodology, the importance of accuracy cannot be understated. This considerable disparity in performance prompts us to better consider ways to enhance the performance of our techniques.

5.2.1 Chain of Thought Analysis

We will now present some examples from responses where our Llama model correctly classified input utterances to their correct intent classes. We will judge these responses by their Chain of Thought, both their presence or absence, as well as the quality of it. For each prompting technique we tested we will once more present 3 example responses, 1 from each run, which when one considers the size of this test dataset here is an even stronger indication as to whether each prompting technique produces Chains of Thoughts, and if yes, it will give us

insights as to their quality. More such example responses can be found in Appendix C, which we also took into account for the inferences made in this subsection.

Deep Breath

text: If the transfer details have already been reviewed and confirmed that they are correct, what other reason would cause my transfer to be declined?

response: sure! based on the text the user wrote, the intent is: "declined_transfer" because the user is asking about reasons why their transfer might be declined, which aligns with the description of the "declined_transfer" intent.

text: How do I get the contactless feature to work for my card?

response: the intent of the user when they wrote the text "how do i get the contactless feature to work for my card?" is "contactless_not_working".

text: Is there any documentation needed for the identity check?

response: the intent of the user when they wrote the text "is there any documentation needed for the identity check?" is "verify_my_identity".

The same issue observed with the technique of "Deep Breath" on the CLINC-150 dataset, where there was a notable lack of Chains of Thought in responses, is also evident on the BANKING77 dataset. In 2 out of the 3 instances, the use of this prompting strategy did not yield a Chain of Thought; the responses primarily consisted of the text utterance paired with a predicted intent class. However, on the 1 occasion where a Chain of Thought was generated, it was accurate and insightful, providing clear reasoning behind the model's choice of intent class. Given that this approach achieved the highest accuracy on this dataset, the scarcity of Chains of Thought, which can also be observed in the additional examples of Appendix C, leaves one desiring more of such detailed processing to truly consider it the best zero-shot prompting method.

Show your Thoughts

text: Can my 19 year old daughter open a savings account at the bank?

response: the intent of the user when they wrote the text "can my 19 year old daughter open a savings account at the bank?" is "age_limit". this is because the user is inquiring about the minimum age requirement for opening a savings account, which falls under the "age_limit" intent category.

text: If I am overseas, how do I change my pin?

response: the intent of the user when they wrote the text "if i am overseas, how do i change my pin?" is "change_pin" because the user is asking about modifying or setting a new personal identification number for their card or account while being abroad.

text: How old do you have to be to be able to open an account?

response: the intent of the user when they wrote the text "there's a direct debit payment in my account that i didn't make" is "direct_debit_payment_not_recognized". this is because the user is indicating that there is a direct debit payment in their account that they did not make, which aligns with the description of the "direct_debit_payment_not_recognized" intent.

In this case, the distribution of texts featuring a Chain of Thought versus those without is inverted as out of 3 responses, all three include a Chain of Thought. In each response the explanation provided by Llama was accurate, offering a logical justification for why the model classified the input utterance into the specified intent class. This is also evident in the examples of Appendix C.

Let's Think

text: Is there a discount for frequently exchanging currencies?

response: the intent of the user when they wrote the text "is there a discount for frequently exchanging currencies?" is "exchange_charge".

text: Why did it decline my payment?

response: the intent of the user when they wrote the text "why did it decline my payment?" is "declined_card_payment". this is because the text explicitly mentions a declined payment, which aligns with the description of the "declined_card_payment" intent.

text: Where will I find my card?

response: the intent of the user when they wrote the text "where will i find my card?" is "order_physical_card". this is because the text is asking about finding a physical card, which aligns with the intent of requesting a new physical card.

The responses generated using the *Let's Think* prompting method entails a generated Chain of Thought in 2 out of 3 cases with the one remaining case consisting of just the text utterance supplied in the prompt and a predicted intent class. the responses that do include a Chain of Thought, are accurate and correct, mirroring the quality seen in previous techniques we've examined. These Chains of Thought are logically sound, albeit straightforward, and not highly complex. The same observations can also be made about the examples found in Appendix C.

Few Shot Experimentation

6.1 Introduction

Upon observing the diminished accuracy of our current prompting techniques (at least when compared to the BERT classifier) as well as the fact that the prompting techniques in both datasets generated Chains of Thought with adequate quality and accuracy, we decided to take our experimentation one step further.

The way we did this was inspired by Wei et al. (2022), where the authors introduced hand-crafted Chains of Thought to the prompt for In-Context Learning. This approach builds on the findings that adding just a few exemplars to a prompt can significantly boost the performance of a Large Language Model, as demonstrated by Brown et al. (2020). One extra benefit of this approach is that the exemplars will guide the model in how to respond to queries. By incorporating exemplars with a generated Chain of Thought in the prompt's context, the model is encouraged to answer similarly.

We created our exemplars for In-Context Learning in a completely different way from the one Wei et al. (2022) utilized since our exemplars for In-Context Learning are not hand-crafted. Instead, they are crafted using the Llama model and the pipeline we described in previous chapters. To do this we followed a very specific procedure, which we will now describe.

Firstly, we extracted the top 5 intent classes for each utterance of the **training** split of each dataset using the same process we described in the previous chapters. We decided to use the prompting technique that gave us the best zero-shot results in each dataset in the few-shot scenario, namely *Show your Thoughts* for CLINC-150 and *Deep Breath* for BANKING77. This was done because, with the In-Context exemplars featuring a Chain of Thought, we expect that the few-shot prompting technique will retain the accuracy advantage of the previous techniques but also enhance the quality of the Chains of Thought that are produced. After shuffling the complete train split of each dataset, and applying the corresponding prompting technique on just the 20 first texts of each training split, we fed those prompts to the Llama model, which generated the 20 most probable results for each prompt utilizing the `top_k` argument of the transformers library's text generation pipeline. After careful examination of these $20 \times 20 = 400$ responses, we chose those that seemed to best fit our criteria to be introduced as exemplars for few-shot prompting.

Those criteria are three:

- The response must have predicted the correct intent class
- The response must include a Chain of Thought
- The Chain of Thought produced must be correct and reasonable

Additionally, we paid close attention to a seemingly minor yet crucial detail: the Chain of Thought must precede the model's predicted intent. This order is vital because if the model first predicts the class and then explains its reasoning, it appears to be justifying its decision. In contrast, we aim for the model to first deliberate on the utterance, present its intermediate reasoning steps, and only then conclude with the class prediction.

If any of the 400 responses fit these three criteria it was immediately considered a candidate response to be included in the context of the prompt for In Context Learning. Out of these candidate responses, we finally kept the 3 responses that we deemed to be the best. Out of these three, we chose one to be included in our 1-shot experiments, two to be included in our 2-shot experiments, and all three were included in our 3-shot experiments.

The prompts that we used for In Context Learning for each dataset, along with full prompt examples can be found in Appendix A.

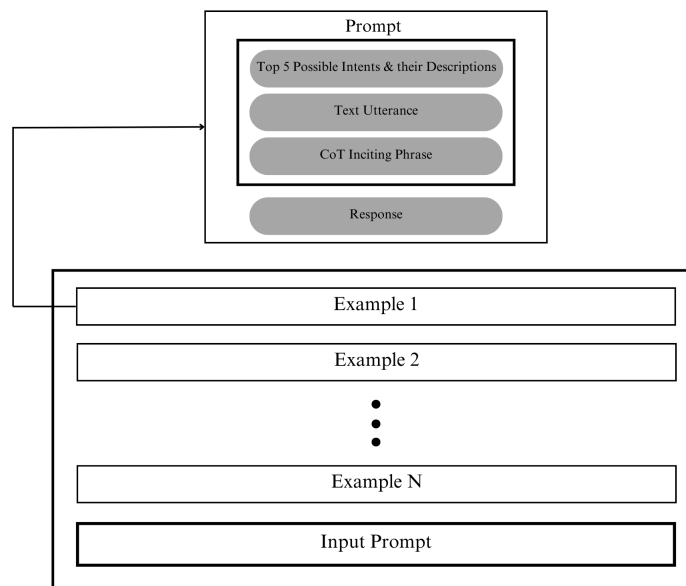


Fig. 6.1: A N-shot prompt and the parts that comprise it.

6.2 CLINC-150

We will now present the few-shot experimentation results using the *Show your Thoughts* prompting technique on CLINC-150, which we presented in Section 3.1, provided us with the best results on the 0-shot experimentation. Recall that CLINC-150 features 150 different intent classes across 10 domains. Our test dataset holds just 5 texts per intent, bringing the total size of our test dataset to 750 texts. This is the same dataset that was used for the 0-shot experimentation.

Prompting Technique	1st run	2nd run	3rd run	Average	St. Dev.
Show your Thoughts	86.00	88.13	86.40	86.84	1.13
1-shot	85.33	88.13	84.40	85.20	1.94
2-shot	88.53	87.33	88.93	88.26	0.83
3-shot	88.53	89.06	89.06	88.88	0.30
BERT	96.00				

Tab. 6.1: Accuracy of CLINC-150 across 3 runs and their average.

The analysis reveals that all three employed techniques consistently yield impressive results, with each maintaining an accuracy rate exceeding 87% in all but two instances, and surpassing 88% in all but three. Notably, the 3-shot technique emerges as the superior method among the three, recording an exceptional average accuracy of 88.88% across three separate runs and a standard deviation of 0.3. Delving into the individual run performances, the initial run registered the lowest accuracy at 88.53%, while both the second and third runs mirrored each other with a slightly higher accuracy of 89.06%. The remarkable aspect of these results lies in the minimal variance in accuracy scores, with a mere 0.53 percentage points range, marking the narrowest margin observed so far in all the 0-shot prompting techniques utilized. This consistency in performance suggests that the 3-shot technique not only demonstrates high accuracy but also exhibits a notable level of reliability and stability, making it a dependable method for applications.

The 2-shot technique emerged as the second most effective achieving an average accuracy of 88.26% across three runs, which is just 0.6 percentage points worse than the 3-shot technique and a standard deviation of 0.83. Notably, its most successful run was the final one, which attained an 88.92% accuracy rate. This score is particularly noteworthy as it stands as the second highest overall when considering the tie for the best score. On the lower end, the least successful run of the 2-shot technique recorded an accuracy of 87.33%. The variability between

the highest and lowest accuracy in these runs was just 1.6 percentage points, a remarkably narrow margin.

In contrast to the other techniques, the 1-shot approach is the least effective, achieving an average accuracy of only 85.2% on the test dataset, which is nearly 3 percentage points below the average accuracy of the 2-shot method and 3.6 percentage points below the average accuracy of the 3-shot method, while attaining a standard deviation of 1.94. The results of the 1-shot technique showed considerable variability, with a significant margin of nearly 4 percentage points between its best and worst outcomes. The best performance of this method was observed in its second run, achieving an impressive 88.13% accuracy, which ranks as one of the best scores across all techniques. However, this high point contrasts sharply with its lowest accuracy of 84.4%, recorded in the third run, while an 85.13% on the 1st run brought the average to the point it stands.

This wide gap in performance levels raises questions about the reliability of the 1-shot technique. The exceptional result in the second run, which is markedly different from the other two runs, might indicate an anomaly or outlier. Such an inconsistency undermines the credibility of the 1-shot method, casting doubt on its dependability. A reliable technique should ideally exhibit consistent and predictable results, which the 1-shot approach seems to lack, as evidenced by the disparate outcomes in its various applications.

The performance of the 1-shot *Show your Thoughts* technique is further undermined, not just by the wide gap between its best and worst accuracy scores, but also by its inability to surpass the average accuracy of any 0-shot prompting technique discussed in the previous chapter. Specifically, the 0-shot *Show your Thoughts* technique achieved a notable average accuracy of 86.84%, while the least effective among the previously mentioned 0-shot prompting techniques, *Let's Think*, recorded an average accuracy of 85.46%. When these figures are compared, it becomes evident that the 1-shot *Show your Thoughts* lags as it scores 1.6 percentage points lower than the 0-shot *Show your Thoughts* and is 0.26 percentage points less accurate than *Let's Think*. This comparative analysis clearly illustrates that the 1-shot version does not measure up to the standards set by the 0-shot techniques. Its lower average accuracy, coupled with its inconsistency (as evidenced by the large variance in its accuracy scores), severely impacts its reliability, leading to the conclusion that it is a less trustworthy method in the context of the techniques being evaluated.

The final stage of our analysis involves a comparative evaluation between the three few-shot prompting techniques and the BERT classifier, which was fine-tuned on this dataset and

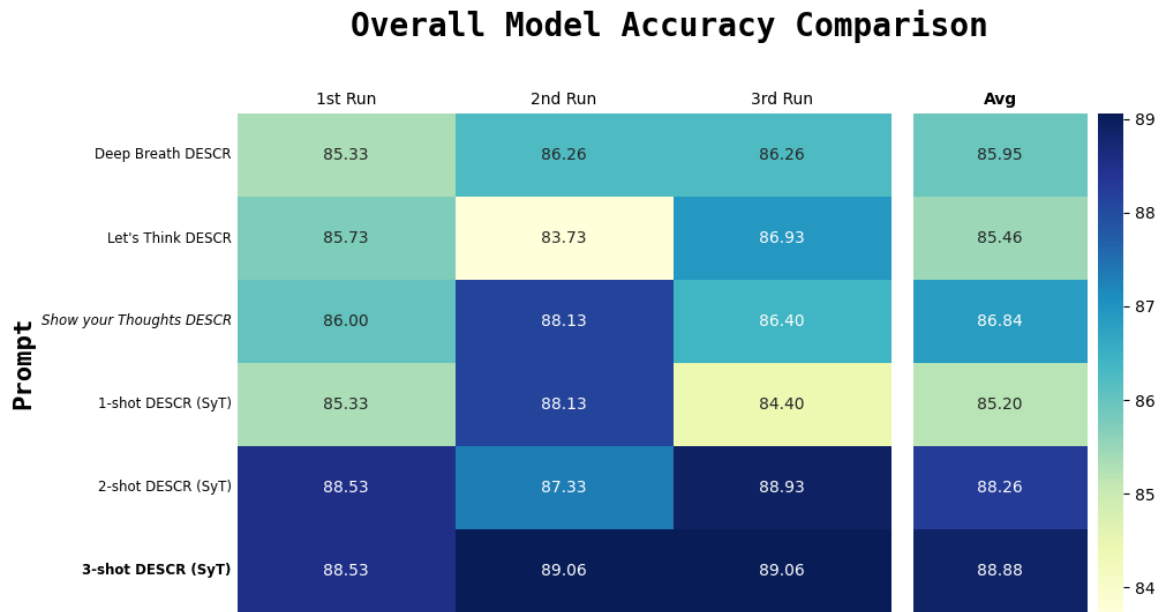


Fig. 6.2: A heatmap of all the results, both individual runs and average accuracy, using the description strategy for CLINC-150. We can easily discern between the best and worst runs and methods with ease and it becomes apparent that the best method in terms of accuracy is the 3-shot method.

utilized specifically for extracting the top 5 intents per text. While our primary focus with the BERT classifier was on its top-5 accuracy, to ensure the most accurate extraction of the top-5 predicted intents, we also kept track of its classic accuracy. On the entire test dataset, the BERT classifier impressively achieved a 96% accuracy rate, while on the same test subset used for our prompting experiments, it again achieved a 96% accuracy.

In this context, the performance of our 3-shot *Show your Thoughts* technique is particularly noteworthy. It managed to approach the high accuracy of the BERT classifier, with its best result hitting 89.06% and an average accuracy of 88.88%. However, it is crucial to emphasize that while BERT significantly outperformed our technique in terms of accuracy, these two models are fundamentally designed for different purposes. BERT is primarily utilized for tasks such as Masked Language Modeling and various classification tasks, while the Llama model we employed is geared toward text generation. The Llama model is not inherently suited for classification tasks, at least not in the manner that BERT is.

Therefore, the observed disparity in their performance results is to be expected given their differing functionalities and design objectives. However, the fact that our 3-shot *Show your Thoughts* technique was able to narrow the gap in performance to some extent with BERT is promising.

6.2.1 Chain of Thought Analysis

After presenting the outcomes of our few-shot experiments, we will proceed to examine 3 randomly selected responses (1 from each run) for each of the 3 few-shot techniques we discussed just before, all of which were accurately classified. We will only present correctly classified examples here because in the following chapter we present a thorough error analysis. The purpose of this examination is to determine whether these responses exhibit a Chain of Thought and to evaluate the quality and veracity of this Chain of Thought. It is important to note that, given the nature of our few-shot experiments and the fact that the in-context examples we employed include responses characterized by a Chain of Thought, we anticipate that nearly all of these responses will demonstrate a Chain of Thought. However, it remains essential for us to verify this expectation through a detailed review.

More response examples can be found in Appendix B which we also took into account for the inferences that we made in this subsection.

1-shot

text: i need to know the carry-on rules for southwest

response: since the user needs to know the carry-on rules for southwest airlines, which falls under the category of seeking information about carry-on luggage, the intent is "carry_on".

text: talk fast

response: since the user requested to talk fast, which is related to changing the speaking speed, the intent is "change_speed".

text: travel alert in cuba

response: since the user is looking for information about travel alerts in cuba, which falls under the category of seeking information about potential disruptions or noteworthy events that may impact travel plans, the intent is "travel_alert".

As anticipated, every response generated by the Llama model showcased a Chain of Thought that it developed. Each response was not only truthful but also followed a logical reasoning path that culminated in the correct answer. Additionally, it's worth noting that the model benefited from the extra care we took in presenting the in-context exemplars. Specifically, we made sure to include the final answer regarding the intent class at the end of each sentence following the Chain of Thought. Since Llama2 is a decoder-only model with masked attention that considers only the previously generated text, if the predicted class is output before the Chain of Thought, the decoder is not aware of the Chain of Thought at the time of making

the prediction. Therefore, the following text could be seen as an after-the-fact explanation of its decision, which cannot be considered a Chain of Thought, as Wei et al., 2022 defined Chain of Thought as a series of *intermediate* reasoning steps. The same observations also apply to the examples of Appendix B.

2-shot

text: how soon milk expires

response: based on the text provided, the user's intent is to seek information about the shelf life or expiration date of milk, which falls under the category of "food_last", as the user is inquiring about the remaining time until the milk goes bad.

text: should i have my shots up to date before i travel to the virgin islands

response: based on the text provided, the user's intent is to seek information about the necessary vaccinations before traveling to the virgin islands, specifically inquiring about whether they should have their shots up to date. therefore, the intent of the user is "vaccines".

text: let me know how many pounds are in 10 kilos

response: based on the text provided, the user's intent is to seek information about the conversion of weight measurements from kilograms (kg) to pounds (lb), specifically to determine how many pounds are equivalent to 10 kilograms. therefore, the intent of the user is "measurement_conversion".

The observations noted for the Chains of Thought created through our 1-shot approach are also applicable here since the responses of our 2-shot method are better articulated, with enhanced clarity, and exhibit more sophisticated reasoning behind the classification of each text utterance into its respective intent class. This development is highly encouraging for our objectives, as it allows us to confidently affirm the efficacy of our methodology. This assertion is further bolstered by the average accuracy rates achieved by the model. The same observations also apply to the examples of Appendix B.

3-shot

text: give me the date that i last got my car checked

response: based on the given text, the intent of the user when they wrote "give me the date that i last got my car checked" is to seek information about the last time their car was serviced or maintained, which aligns with the "last_maintenance" intent from the provided set of intents.

text: how many stars does brasserie les halles have

response: based on the given text, the intent of the user when they wrote "how many stars does brasserie les halles have" is to seek information about the rating or reputation of brasserie les halles, which aligns with the "restaurant_reviews" intent from the provided set of intents.

text: i am asking you to add this song to my country music playlist

response: based on the given text, the intent of the user when they wrote "i am asking you to add this song to my country music playlist" is to update their existing music playlist by adding a new song, which aligns with the "update_playlist" intent from the provided set of intents.

As anticipated, the 3 sample responses generated using our 3-shot technique exhibit Chains of Thought that are, not only more complex but also more eloquently articulated than those produced by our 2-shot method. This observation serves as strong evidence of the effectiveness of our approach, especially when applied to this particular dataset. It demonstrates the capacity of our method to enhance the quality and sophistication of the generated responses, thereby underscoring its potential as a robust tool for improving the interpretability and accuracy of response generation in complex datasets. The same observations also apply to the examples of Appendix B.

6.3 BANKING77

We now present the findings of our few-shot experiments, which employed the *Deep Breath* prompting technique. Notably, this technique emerged as the most effective in our prior 0-shot experimentation using the same dataset. We chose this prompting technique for our few-shot experimentation because it was the one that gave us the best 0-shot results in terms of accuracy, even though most of the responses did not feature a Chain of Thought. But some of them did feature a Chain of Thought, and through the pipeline we described at the beginning of this chapter we managed to find some exemplars we deemed adequate to be included in our few-shot experiments.

Prompting Technique	1st run	2nd run	3rd run	Average	St. Dev.
Deep Breath	74.54	72.20	73.50	73.31	1.17
1-shot	71.68	69.35	72.98	71.32	1.83
2-shot	70.90	71.42	72.46	71.59	0.79
3-shot	73.50	72.98	74.54	73.67	0.79
BERT	94.00				

Tab. 6.2: Accuracy of BANKING77 across 3 runs and their average.

Upon initial examination of the few-shot results, we observe a pattern that mirrors that of the few-shot experiments conducted on CLINC-150. Once again, the 3-shot technique emerges as the top performer with an average accuracy of 73.67% and a standard deviation of 0.79. This performance not only stands out on this specific dataset but also exceeds the score of the 0-shot *Deep Breath* method, which recorded an average accuracy of 73.31%. Although the difference in accuracy between the 0-shot and 3-shot *Deep Breath* techniques might appear minimal by CLINC-150 standards, it's important to acknowledge the increased complexity of this dataset. Designed to more closely mimic real-world input utterances, this dataset presents challenges not encountered in CLINC-150.

Indeed, as previously noted, the 3-shot technique emerged as the most effective in our few-shot experimentation, registering an average accuracy of 73.67%. Delving deeper into the individual runs of the 3-shot experiments, we observe notable variation in performance. The third run outperformed the others, achieving an accuracy of 74.54%. This performance is particularly significant as it represents the best performance among all nine runs of the few-shot experiments and is tied for the best result overall with the first run of 0-shot *Deep Breath*. On the other end of the spectrum, the second run of the 3-shot experiments attained an overall accuracy of 72.98%, a score which falls while the first run occupied a middle ground relative to the other two, recording a score of 73.5%.

The variability in accuracy across these runs - approximately 1.56 percentage points - is the lowest observed thus far on this dataset. This shows that maybe adding in-shot exemplars to this particular prompting technique, i.e., *Deep Breath*, could help stabilize the results a bit, as 0-shot *Deep Breath* had a margin of 2.34 percentage points between the best and worst results.

The 2-shot technique emerged as the runner-up in terms of effectiveness, achieving an overall accuracy of 71.59%, which is 2.08 percentage points below the accuracy of the 3-shot and, quite notably, 1.72 percentage points below the average accuracy of *Deep Breath*, which as we mentioned before achieved an average accuracy of 73.31%, while the standard deviation came to 0.79. It also scored worse than *Show your Thoughts* which was the second best 0-shot method, and only merely surpassed *Let's Think*, the worst of our 0-shot methods. It also achieved a 0.79 standard deviation, which is the same as the standard deviation of the 3-shot method. Within this series of runs, the most best was observed in the third attempt, registering an accuracy of 72.46%, a result that as we can see on the table is pretty average when compared with all the other runs. The initial run of the 2-shot technique yielded a moderate 70.9% which was the least impressive run of the 2-shot method and one of the lowest scores we observed

either in a few-shot or 0-shot setting. The performance of the second run in the 2-shot setting is right in the middle between the first and third runs, achieving an accuracy of 71.42%. The margin between the best and worst results on this series of experiments comes to just 1.56 percentage points which is tied with the 3-shot margin for the lowest margins observed in this particular dataset. This comes as a stronger signal that the extra in-shot exemplars help stabilize the results and make our model more sure of its responses, no matter the fact that these responses could be wrong a little more often in this 2-shot setting when compared with the 0-shot and 3-shot variants of *Deep Breath*.

The worst of these few-shot techniques was, like it was on CLINC-150, the 1-shot method, which performed very similarly to the 2-shot method, and more specifically a 71.32% which is just 0.27 percentage points behind the 2-shot method and 2.35 percentage points behind the 3-shot method, with a standard deviation of 1.83. This method, as the 2-shot method did, scored worse than every 0-shot method we experimented with as the lowest of the 0-shot scores was 71.33%, just 0.02 percentage points over 1-shot, but still over 1-shot. The best run of the 1-shot experiments was the 3rd one which achieved a pretty good 72.98% accuracy, resembling the results of the 0-shot *Deep Breath* experiments, while the worst run was the second one, managing only a 69.35% overall accuracy, while the 1st run sits somewhere in the middle of the previous two scores, at 71.68%. As we can already see, the variability between the best and worst runs is around 3.6 percentage points which is pretty substantial, even for this dataset which always gave us results with high variability in their accuracy between runs.

The dataset in question, as pointed out by Casanueva et al. (2020), is designed to closely mirror real-world data, making the task of achieving high-performance scores more challenging. This inherent difficulty could be a primary factor contributing to the observed performance levels, and might also explain why the technique of few-shot prompting did not yield as significant improvements here as it did with the CLINC-150 dataset. Another aspect to consider is the specific in-shot exemplars selected for use with this dataset which even though they fulfilled the predetermined criteria as to their quality, the criteria themselves might not be adequate to help us choose exemplars that will yield a significant improvement in performance on this particular dataset. Additionally, the limited number of exemplars used in this study, which was restricted to just three, could also be a contributing factor to the sub-optimal performance observed. This raises the question of whether increasing the number of exemplars might have produced different results.

Another perspective on the observed performance levels could be attributed to the specific variant of Llama utilized in this research, namely Llama2-13B. Considering the scale of language

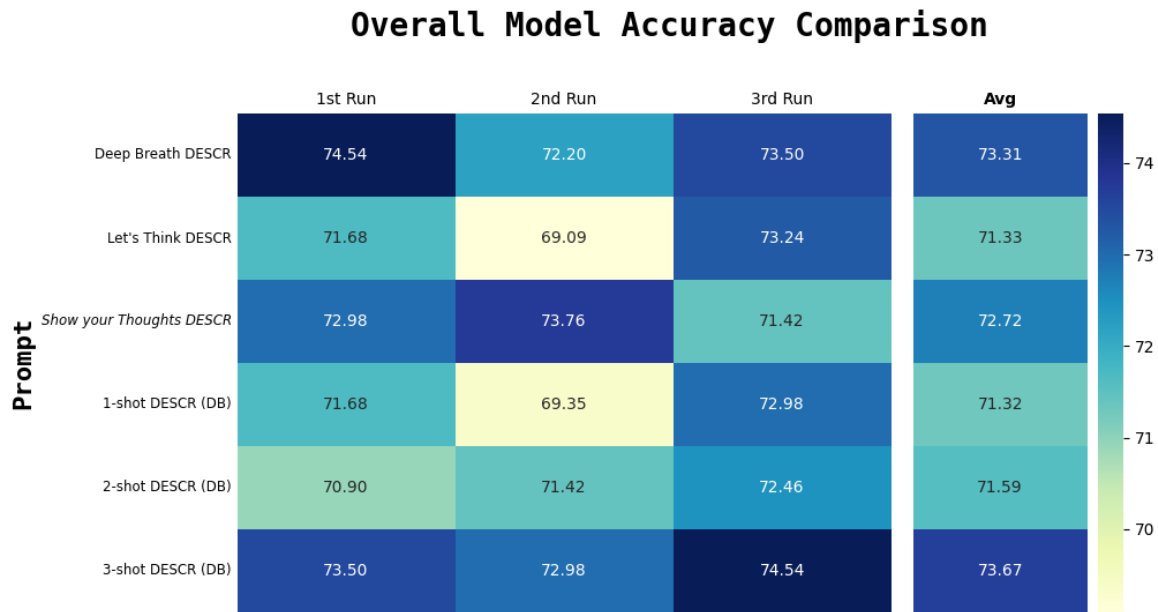


Fig. 6.3: A heatmap of all the results, both individual runs and average accuracy, using the description strategy for BANKING77. We can easily discern between the best and worst runs and methods with ease and it becomes apparent that the best method in terms of accuracy is the 3-shot method, even by a very slight margin.

models, a 13 billion parameter model may not possess sufficient capacity for this dataset. This viewpoint gains credence when reflecting on the work of Loukas et al. (2023), who conducted a similar classification task employing larger language models, including GPT-4, which, while its exact parameter count remains undisclosed, is speculated to be in the hundreds of billions. Notably, this effort was undertaken without leveraging techniques such as Chain of Thought or providing descriptions, relying solely on In Context exemplars from each class. Despite the significantly larger scale of GPT-4, the improvement in results over those achieved with Llama2-13B was not markedly superior. This observation suggests that while the size of a language model is a critical factor, the application of more sophisticated prompting techniques or the challenges of the dataset may limit the extent of performance gains achievable solely through increased model capacity.

The comparison of these outcomes with the findings from the fine-tuned BERT classifier, which was employed to extract the top-5 potential intents for each utterance, further underscores the inadequacies observed. BERT achieved an impressive overall accuracy of 92% on the entire test dataset and 94% on the test set that was used for inference with Llama. In contrast, our best performing few-shot method, the 3-shot technique, only managed an average accuracy of 73.67%, which is almost 20% lower than the accuracy achieved by the BERT classifier, 18.33%

to be exact. This performance manages to close the gap between the BERT-classifier and the Llama-based approach to this classification task, even by a very small margin.

This contrast underscores a crucial point: while BERT may be from an earlier generation of models and is designed primarily for tasks like classification, our Llama model, aimed at a different function as a text generation model, faces distinct challenges. It becomes clear that to even approach the performance level of previous generation models like BERT, our Llama model requires extremely careful handling and a highly specialized approach, underlining the significant differences in model capabilities and design objectives between the two.

6.3.1 Chain of Thought Analysis

After presenting the results of our few-shot experiments, we will move on to analyze 3 randomly selected responses (1 from each trial) for each of the three few-shot techniques we previously discussed, all of which were accurately classified. This analysis aims to check if these responses exhibit a Chain of Thought and to assess the quality and accuracy of this Chain of Thought. Considering the nature of our few-shot experiments and the fact that the in-context examples we used are marked by a Chain of Thought, we expect almost all of these responses to display a Chain of Thought. Nonetheless, we must confirm this expectation by conducting a thorough examination.

More such examples can be found in Appendix C, which we also took into account for the inferences made in this subsection.

1-shot

text: Where can i find your locations?

response: based on the text "where can i find your locations?", the user is seeking information about the physical locations where the services or products offered by the entity are available, and therefore the intent of the user is "country_support".

text: Does delivery to the us take long?

response: based on the text "does delivery to the us take long?", the user is inquiring about the estimated delivery time for a card to the united states, specifically asking about the duration of the delivery process. therefore, the intent of the user is "card_delivery_estimate".

text: How come the atm gave me less cash than what i asked for?

response: based on the text "how come the atm gave me less cash than what i asked for?", the user is seeking clarification and understanding regarding a discrepancy between the amount of money requested and the amount dispensed by an automated teller machine (atm), which aligns with the intent "wrong_amount_of_cash_received".

The responses generated by our model using the 1-shot technique demonstrate eloquent and precise Chains of Thought for every piece of text, aligning with our expectations due to the inclusion of an exemplar response featuring a Chain of Thought in the 1-shot prompt. Additionally, a noteworthy aspect of these responses is that the predicted intent is positioned at the end, allowing the Large Language Model to first construct a reasoning path before reaching the desired conclusion. The same observations apply to the examples of Appendix C.

2-shot

text: Where's the verification code for the top-up card?

response: based on the text "where's the verification code for the top-up card?", the user is seeking confirmation of the addition of funds to their account or service, specifically asking for the verification code related to the top-up card. therefore, the intent of the user is "verify_top_up".

text: Where do i go to get the code to verify the top up card?

response: based on the text "where do i go to get the code to verify the top up card?", the user is seeking information on how to obtain a verification code to confirm the addition of funds to their account. therefore, the intent of the user is "verify_top_up".

text: My card is about to expire,

response: based on the text "my card is about to expire," the user is seeking information or assistance regarding the upcoming expiration of their payment card. therefore, the intent of the user is "card_about_to_expire."

Similarly to what we have observed across all responses generated by the few-shot methods on both datasets, the responses from the 2-shot method consistently display correct and truthful Chains of Thought that are articulated eloquently. Here as well, we note that the reasoning precedes the predicted intent, which aligns with our objective and was a key factor in our search for suitable in-shot exemplars. This deliberate structuring to prioritize reasoning before concluding is a testament to the effectiveness of our approach in crafting responses that meet our desired criteria. The same observations apply to the examples of Appendix C.

3-shot

text: Can I cancel a transfer if is already processed?

response: the user is asking if they can cancel a transfer that has already been processed, which falls under the category of attempting to halt or reverse a previously initiated monetary transaction. therefore, the intent of the user is "cancel_transfer".

text: I think something may have happened that caused a charge to show up twice.

response: based on the text "i think something may have happened that caused a charge to show up twice," the user is seeking clarification and resolution regarding a double charge on their account. therefore, the intent of the user is "transaction_charged_twice."

text: Is it possible to make several disposable cards in a day?

response: the user is asking if it is possible to create multiple disposable cards in a day, which falls under the category of "disposable_card_limits." therefore, the intent of the user is "disposable_card_limits."

In this instance, we observe that, in some cases, the model explicitly mentions the intent class name before providing its final answer, a pattern not seen in the responses from the two previous few-shot sets. This difference may be attributed to the inclusion of a third in-shot exemplar, which was absent in the previous few-shot experiments, or to its interaction with the other two in-shot exemplars. Regardless, we consider this to be a minor observation and choose to overlook it, given that the resulting Chains of Thought are well-articulated and truthful across all 3 responses. The same observations apply to the examples of Appendix C.

Ablation Study & Error Analysis

7

7.1 Ablation Study

Our next step involves conducting an ablation study, a crucial part of our analysis. The primary objective of this study is to evaluate the robustness and efficacy of our technique. Specifically, we aim to determine the impact of incorporating intent descriptions and the inclusion of Chain of Thought phrases. This will help us understand if these elements significantly enhance our results, contributing to a deeper understanding of the elements that drive the success of our methodology.

To conduct this comprehensive ablation study, our approach will be methodical and staged. Initially, we will modify the prompt by omitting the Chain of Thought component. In this phase, our task for Llama will be straightforward: to classify the provided text utterance into one of the top-5 intent classes we have predefined. Importantly, during this initial step, we will retain the intent descriptions in the prompt to isolate the effect of removing only the Chain of Thought.

Following this, we will invert our approach: we will preserve the Chain of Thought portion of the prompt but will eliminate the intent descriptions. This will enable us to specifically gauge the influence of removing the intent descriptions on the accuracy of our classification task.

The final segment of our ablation study aims to assess the combined importance of both techniques. We will do this by evaluating the accuracy of a prompting technique that excludes both the Chain of Thought and the intent descriptions. This will provide us with valuable insights into the individual and collective contributions of these elements to the overall effectiveness of our methodology.

For the reader to accurately discern between the methods that include Descriptions and those that do not, we will from now on denote the techniques that do feature descriptions with their name and the **DESCR** concatenated.

7.1.1 Without Chain of Thought

For this initial part of our ablation study we directly query our Llama model to choose the most accurate intent class from the five we presented to it presented along with their descriptions. The way we do this is by simply asking Llama: *What is the intent of this text: [x]* where [x] is the utterance of the dataset.

One example of a full prompt utilizing this method is this:

Question: We have the following set of intents along with their descriptions:

- * **no**: The intent “no” signifies a straightforward refusal or a negative response to a question or statement, indicating a lack of agreement or affirmation.
- * **cancel**: The intent “cancel” involves users expressing a desire to cancel or terminate a particular action or process.
- * **repeat**: The intent “repeat” involves asking for the duplication or reiteration of a statement, command, or information previously provided, seeking clarification or emphasis.
- * **who_do_you_work_for**: The intent “who_do_you_work_for” involves inquiries seeking information about the entity or organization that an individual or system is affiliated with or employed by.
- * **change_speed**: The intent “change_speed” involves seeking information or guidance on adjusting the speed settings, such as playback speed or talking speed, on a device or application.

What is the intent of this text: ”forget what i told you to do”?

Answer: [intent name goes here].

Let us now present the results of this technique for both datasets.

Prompting Technique	1st run	2nd run	3rd run	Average	St. Dev.
No-CoT DESCR	86.53	85.86	87.46	86.61	0.80
Deep Breath DESCR	85.33	86.26	86.26	85.95	0.54
Let's Think DESCR	85.73	83.73	86.93	85.46	1.61
Show your Thoughts DESCR	86.00	88.13	86.40	86.84	1.13
BERT	96.00				

Tab. 7.1: Accuracy of CLINC-150 across 3 runs and their average. The average accuracy of the method without a Chain of Thought is highlighted.

Prompting Technique	1st run	2nd run	3rd run	Average	St. Dev.
No-Cot DESCR	71.94	73.50	74.02	73.15	1.08
Deep Breath DESCR	74.54	72.20	73.50	73.31	1.17
Let's Think DESCR	71.68	69.09	73.24	71.33	2.09
Show your Thoughts DESCR	72.98	73.76	71.42	72.72	1.19
BERT	94.00				

Tab. 7.2: Accuracy of BANKING77 across 3 runs and their average. The average accuracy of the method without a Chain of Thought is highlighted.

As we can see from the tables above, in both datasets the average accuracy of our latest prompting technique, No CoT, which features the descriptions in the prompt but not a phrase that generates a Chain of Thought, lags behind the best prompting technique by a very narrow margin. No-CoT produced an average accuracy of 86.61% surpassing both *Let's Think DESCR* and *Deep Breath DESCR* by 1.15 percentage points and 0.66 percentage points respectively, being just 0.23 percentage points behind *Show your Thoughts DESCR*. It also produced a standard deviation of 0.8. On BANKING77 the same method produced an average accuracy of 73.15% surpassing both *Let's Think DESCR* and *Show your Thoughts DESCR* by 1.82 percentage points and 0.43 percentage points respectively, while it sits just 0.16 percentage points behind *Deep Breath DESCR* which produced the best results on this Dataset, while it also produced a standard deviation of 1.08.

In CLINC-150 the best result of No-CoT DESCR was an excellent 87.46% on its third run, while the worst run was an 85.85% and the first run was an elevated 86.53%. The margin between the best and the worst results was 1.6 percentage points which is one of the lowest recorded in the CLINC-150 dataset.

In BANKING77 we see a very similar image. The third run was the highest scoring among the three, achieving a 74.02% accuracy, which is the second highest of all runs on this dataset. The first and second runs procured a 71.94% and 73.5% respectively, bringing the variability

between the best and worst scores to 2.08 percentage points, the lowest observed in this dataset.

The analysis indicates that in both datasets evaluated, the No-Cot DESCOR method, while not emerging as the top performer in the 0-shot setting, was remarkably close to the leading methods in terms of performance. This observation leads to an intriguing conclusion: even though prompts utilizing the Chain of Thought approach were the frontrunners in both scenarios, the absence of Chain of Thought prompting would likely not result in a significant disparity in accuracy. This inference is in partial alignment with the research findings presented by Kojima et al. (2022), Wei et al. (2022), Wang et al. (2023), and Wang et al. (2023). These studies have suggested that the accuracy of smaller models (i.e. with less than 100 Billion parameters) on similar tasks to ours could potentially diminish when Chain of Thought prompting is employed. However, it's important to note that this agreement is only partial, as in both datasets analyzed, the methods that secured the top performance spots were those that incorporated Chain of Thought prompting, albeit by a slim margin.

7.1.2 Without Descriptions

We will now test the importance of intent descriptions in the prompt. The way we will do this is by keeping everything else in the prompt the same, but simply not including the descriptions of the top-5 intents which our model will have to choose from. This way the model will have to find similarities between the utterance we provide it and the intents by using just the intent names, which sometimes may be uninformative, or confusing as to what exactly they represent. As for our prompting methods, in this step, we will research every prompting method we have mentioned in previous chapters that is expected to produce a Chain of Thought, both 0-shot and few-shot. For the few-shot scenarios, we opted to test the same method we used on the few-shot scenarios with the intent descriptions present in the prompt, meaning *Show your Thoughts* for CLINC-150 and *Deep Breath* for BANKING77. The exemplars that were used for In Context Learning were chosen using the same methodology we used for the few-shot scenario with the descriptions included.

CLINC-150

After just one look at the table of the results, we notice that the absence of the descriptions has a big impact on the results. The average accuracy of every method we tested, both zero-shot and few-shot, has dropped significantly. When compared with the inclusion of descriptions,

Method	1st run	2nd run	3rd run	Average	St. Dev.	DESCR Average
Deep Breath	74.66	74.13	74.53	74.44	0.27	85.95
Let's Think	78.00	78.26	76.53	77.52	0.93	85.46
Show your Thoughts	76.93	75.86	76.93	76.57	0.62	86.84
1-shot	80.00	79.33	80.66	79.99	0.66	85.20
2-shot	83.33	83.06	85.20	83.86	1.16	88.26
3-shot	82.66	82.66	81.46	82.26	0.69	88.88
BERT	96.00					

Tab. 7.3: Accuracy of CLINC-150 without the use of intent descriptions. The average accuracy of the best method without intent descriptions is highlighted.

some methods' average accuracy dropped a little less, like the 2-shot technique where we notice only a 4.4 percentage points drop in average accuracy, or a lot more, like *Deep Breath* which dropped a whopping 11.5 percentage points in accuracy score.

From the 0-shot techniques the best results now came from *Let's Think* with a 77.52% average accuracy, a standard deviation of 0.93 and variability between best and worse runs of under 2 percentage points which is remarkable when one considers the fact with the descriptions present this same prompting technique had more than 3 percentage points variability. Notably, the drop in performance when compared with the descriptions included is around 8 percentage points, one of the lowest we found.

Second best among the 0-shot techniques came our *Show your Thoughts* method with an average accuracy of 76.57%, a little less than 1 percentage points off the *Let's Think* variant, and 10.27 percentage points below the average accuracy of the same technique when used with the intent descriptions in the prompt. The standard deviation came to 0.62 while the margin here between the best and worst runs sits at an impressive 1.07 percentage points which is less than half of the corresponding margin with descriptions included.

The worst performer for the 0-shot techniques was *Deep Breath*, managing only a 74.44% average accuracy across the three runs and a standard deviation of 0.27. The average accuracy of 74.44% is 2.13 percentage points below *Show your Thoughts* and 3.08 percentage points below *Let's Think*. What is notable though is the drop in performance when compared with the description-rich variant, *Deep Breath DESCR*, as the drop rose to 11.5 percentage points, the largest of the three. As for the margin between best and worst runs, it was just 0.53 percentage points which is the smallest margin we've observed in any 0-shot technique, whether with descriptions or without.

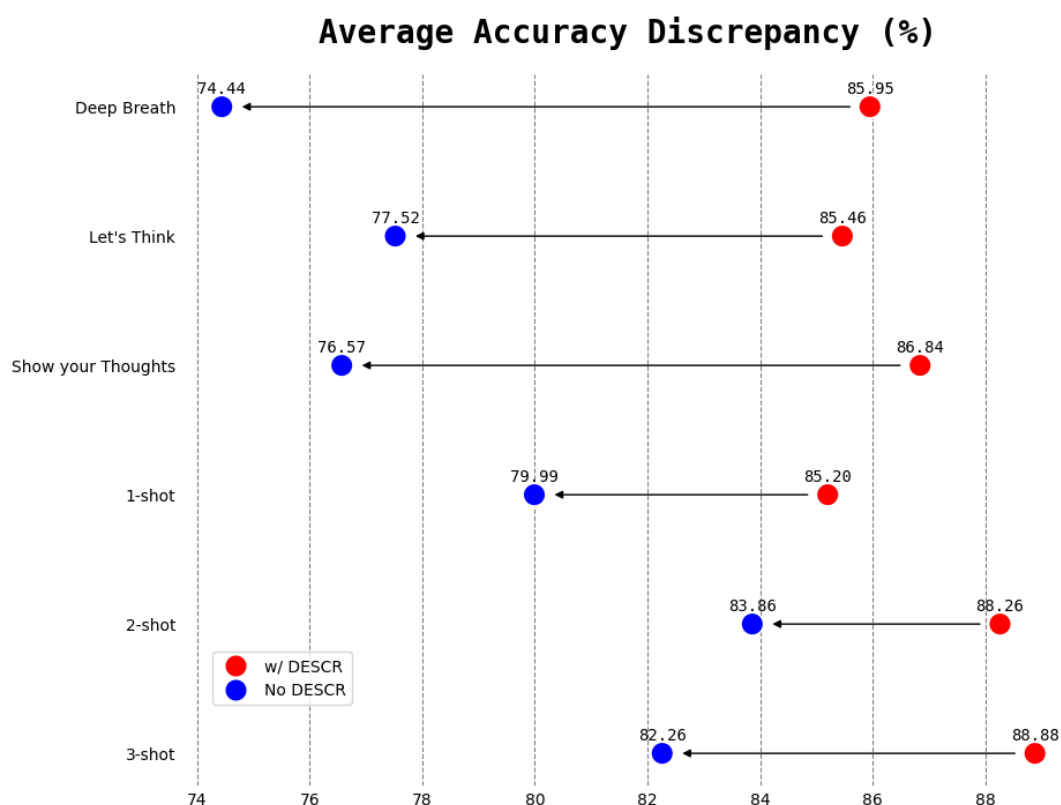


Fig. 7.1: The difference in average accuracy for CLINC-150 between using the descriptions and not using them. It is evident that the technique that profits the most by the intent descriptions is *Deep Breath*, while the few-shot methods seem to not suffer as much.

The few-shot techniques performed slightly better than the 0-shot techniques, as with just 1 in-shot exemplar we managed to get a 79.99% in average accuracy which is 3.42 percentage points larger than the respective 0-shot score and 5.21 percentage points down from the original 1-shot with the descriptions in the prompt. The standard deviation of this technique is 0.66. We also noticed a general stability in the results of 1-shot with the range between best and worst average accuracies sitting a 1.33 percentage points, as the worst score was achieved on the 2nd run, a 79.33%, while the best at the third run, a 80.66%. The first run achieved a solid 80%.

2-shot was the best performer in these few-shot prompting techniques without the intent descriptions as it achieved a very respectable 83.86%, close to 4 percentage points up from the previous technique and just over 4 percentage points below the 2-shot technique with descriptions added, with a standard deviation of 1.16. The best run was the third one with a surprising 85.2%, while the first and second ones performed similarly to one another, with

83.33% and 83.06% respectively. This brings the total variability of the results to 2.14 percentage points which is average for this dataset.

Finally, the performance of 3-shot sat in the middle of the previous two techniques we presented, with an 82.26% average accuracy, 2.27 percentage points over 1-shot, and just 1.4 percentage points below 2-shot. This performance was also 6.62 percentage points worse than the 3-shot performance with the descriptions in the prompt, while achieving a 0.69 standard deviation. Here, the best score was achieved twice, both by the first and second run, which both scored an 82.66% average accuracy, with the third run managing a 81.46% to bring the total variability of the results to 1.2 percentage points, a margin which is good when compared with most techniques.

From the above results, we can easily deduce that descriptions play a very important part in the overall improvement in performance, both in individual runs and when averaged, as not a single prompting technique without descriptions, either 0-shot or few-shot, got anywhere near in terms of accuracy with the accuracy of the respective method with descriptions added. Even the best performer without the descriptions, which was 2-shot with 83.86% average accuracy, was close to 2 percentage points below the worst performer with the descriptions added, i.e. *Let's Think*. This shows that the inclusion of descriptions in the prompt has a very significant impact on the performance of these models.

BANKING77

Method	1st run	2nd run	3rd run	Average	St. Dev.	DESCR Average
Deep Breath	62.33	63.11	63.89	63.11	0.78	73.31
Let's Think	66.75	67.01	64.93	66.23	1.13	71.33
Show your Thoughts	63.89	63.37	66.75	64.67	1.82	72.72
1-shot	64.15	64.93	64.41	64.49	0.40	71.32
2-shot	68.05	68.57	70.64	69.08	1.37	71.59
3-shot	67.27	69.87	68.05	68.39	1.33	73.67
BERT	94.00					

Tab. 7.4: Accuracy of BANKING77 without the use of intent descriptions. The average accuracy of the best method without intent descriptions is highlighted.

From a first look at the table, we immediately notice that descriptions have a similar impact on the performance of BANKING77 too. This can be derived from the fact that the average accuracy of every single technique, both 0-shot or few-shot, drops, and most times in a pretty significant manner. Let us explore these results more deeply.

First and foremost, once more, the best technique without descriptions is *Let's Think*, which was the worst 0-shot performer with the descriptions added to this dataset too. It performed a 66.23% average accuracy which is 5.1 percentage points below the 71.33% this technique scored with the descriptions while attaining a 1.13 standard deviation. The best run was the the second one, which gave us 67.01% while the first and third ones sit at 66.75% and 64.93% respectively. This brings the variability between best and worst results to 2.08 percentage points, which is one of the lowest differences we've observed in this dataset.

The second best among the 0-shot prompting techniques was *Show your Thoughts* with a 64.66% average accuracy and a standard deviation of 1.82. The average accuracy is 8.05 percentage points worse than the corresponding performance with the descriptions added in the prompt, and just 1.56 percentage points below the best performer without descriptions, *Let's Think*. This time around, the best run was the third one with a 66.75%. The first and second runs performed very similarly, with performances of 63.89% and 63.37% respectively. This brings the margin between the best and worst scores to 3.38 percentage points, one of the largest such margins we've seen so far on this dataset.

The worst 0-shot performer this time around was the one that had the best performance with the descriptions added, *Deep Breath*, which scored just a 63.11% average accuracy with a standard deviation of 0.78. This performance is a little more than 10 percentage points below the corresponding performance with the descriptions added, 10.2 percentage points to be exact, while it also lags behind *Show your Thoughts* and *Let's Think* by 1.56 percentage points and 3.21 percentage points respectively. The best run of this technique was the third one with a 63.89% while the first and second ones managed only a 62.33% and 63.11% respectively. The variability between the best and worst scores now sits at 1.56 percentage points, which is the lowest such variability we've seen on this dataset, either with 0-shot or few-shot prompting techniques.

We will now present the BANKING 77 few-shot results of this part of our ablation study. Recall that for these few-shot experiments, we used the *Deep Breath* prompting technique so that we have comparable results for our few-shot techniques both with and without descriptions.

Starting with 1-shot, the average accuracy was 64.49% with a standard deviation of 0.4. The accuracy performance of 64.49% sits not only 6.83 percentage points below the 1-shot technique with descriptions in the prompt but also 1.74 percentage points below the best 0-shot performer without descriptions, *Let's Think*, 0.18 percentage points below *Show your Thoughts* without descriptions, only slightly improving its base method, *Deep Breath*, by 1.38 percentage points.

As for the specific runs, the best one was the second run with 64.93%, while the first and third runs achieved just 64.15% and 64.41%, bringing the margin between the best and worst runs to just 0.78 percentage points which is the lowest margin we've seen on this dataset in any of the prompting techniques, showing us that while not achieving the best results, it's one of the most stable methods of the bunch.

2-shot significantly improves on the 1-shot results, achieving an overall accuracy of 69.08%, which this time surpasses all 0-shot methods by at least 2.9 percentage points, is also 4.58 percentage points better than the 1-shot and 2.51 percentage points below the respective 2-shot method with the descriptions included while the standard deviation is 1.37. The best individual run was the third one with a notable 70.64% while the first and second ones maintained a relatively close 68.05% and 68.57% respectively. This brought the final variability between the best and words runs to 2.59 percentage points, which is just about average for this dataset.

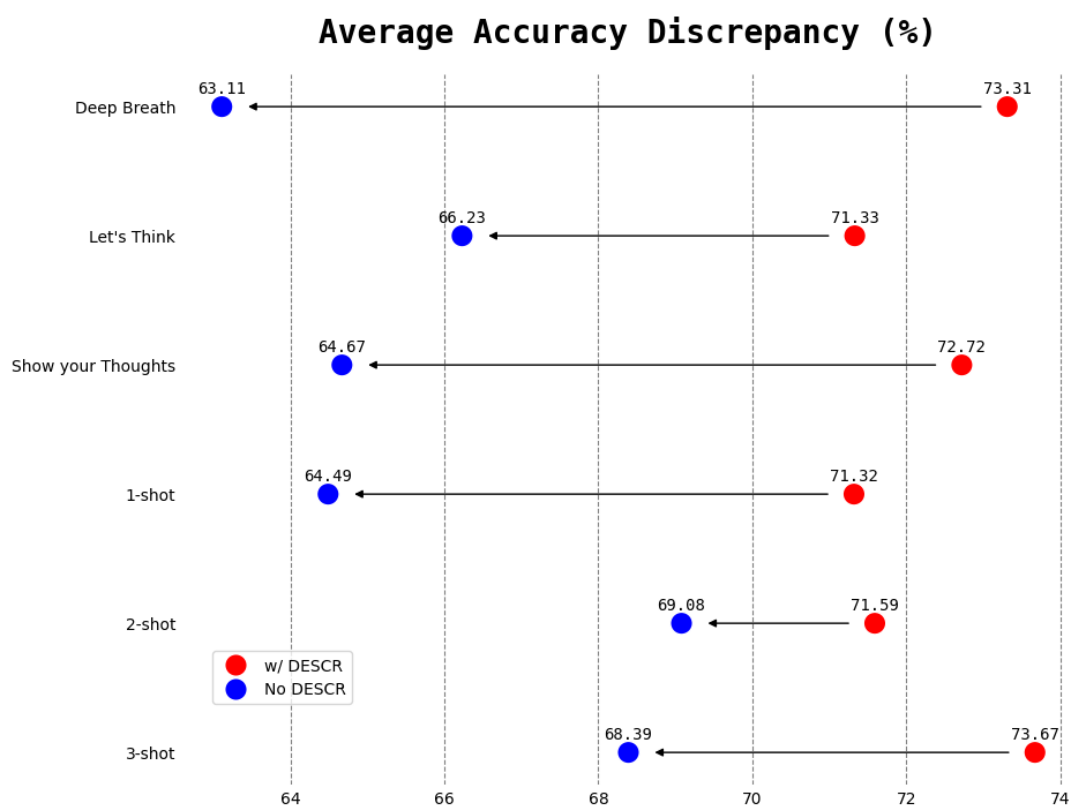


Fig. 7.2: The difference in average accuracy for BANKING77 between using the descriptions and not using them. It is evident that the technique that profits the most by the intent descriptions is once more *Deep Breath*, while the few-shot methods seem to not hurt as much.

Finally, 3-shot could not improve on the 2-shot performance, managing only 68.38% which is 0.69 percentage points below 2-shot, but still 3.72 percentage points better than 1-shot. Also,

this percentage is 5.28 percentage points below the respective method with the descriptions included, which gave us the best results with the descriptions included. The standard deviation of this method is 1.33. The best individual performance was a 69.87% on the second run while the first and third ones achieved a 67.27% and 68.05% respectively, bringing the margin between the best and worst scores to 2.6 percentage points, a pretty average margin for this dataset.

As we can see from these previous results, descriptions play a pivotal role in the improvement of performance in this dataset too. More specifically, every method without descriptions could not surpass any single method with descriptions, even by a relatively small margin on the 2-shot case, though this could and should be attributed to the worsening performance of the 1-shot and 2-shot cases when the descriptions are added to the prompt, where the in-context exemplars lead to a drop in performance.

7.1.3 Without both

For our final test in this ablation study, we will test the case where both the Chain of Thought inciting phrases and the intent descriptions are missing. The way we do this is by simply incorporating into the prompt the top 5 intents as they were extracted using our BERT classifier and simply asking our Llama model to choose the correct one among the 5. The way we question Llama is the same question as in our first part of this ablation study: *What is the intent of this text: [x]* where [x] is the utterance of the dataset.

CLINC-150

Method	1st run	2nd run	3rd run	Average	St. Dev.	DESCR Average
No CoT	80.40	78.66	79.46	79.50	0.87	86.61
Deep Breath	74.66	74.13	74.53	74.44	0.27	85.95
Let's Think	78.00	78.26	76.53	77.52	0.93	85.46
Show your Thoughts	76.93	75.86	76.93	76.57	0.62	86.84
1-shot	80.00	79.33	80.66	79.99	0.66	85.20
2-shot	83.33	83.06	85.20	83.86	1.16	88.26
3-shot	82.66	82.66	81.46	82.26	0.69	88.88
BERT	96.00					

Tab. 7.5: Accuracy of CLINC-150 without the use of intent descriptions, *No CoT* included (highlighted).

We immediately notice that with both the Chain of Thought inciting phrases and the descriptions absent the results are much better than the corresponding ones with just the Chain of Thought inciting phrases added. More specifically, our *No Cot* method achieved an average

accuracy of 79.5% across the three runs with a standard deviation of 0.87. The average accuracy is 1.98 percentage points better than *Let’s Think*, which was the best prompting technique without descriptions, 2.93 percentage points better than *Show your Thoughts*, and an astonishing 5.06 percentage points better than *Deep Breath*. These results seem to once more confirm the findings of Wei et al. (2022), Kojima et al. (2022), Wang et al. (2023), Wang et al. (2023) and others, who supported that performance of Large Language Models with less than 100 billion parameters can be worsened by the use of Chain of Thought prompting. When compared with the few-shot methods without descriptions, it performs slightly worse. To be specific, it performed 0.49 percentage points worse than 1-shot, 4.36 percentage points below 2-shot, and 2.76 percentage points worse than 3-shot. The drop in performance when compared with the description-rich counterpart of this method was also significant though, 7.11 percentage points to be exact. This once more shows that descriptions are a powerful asset to the performance of our prompting techniques and are now proven invaluable.

As to the individual runs of this method, the best one was the first one with an 80.4% overall accuracy, while the second and third ones achieved a 78.66% and 79.46% accuracy respectively. This brings the total variability between the best and worst results to 1.84 percentage points which is just about average for this dataset. This shows that while not the best results, this method is stable between runs.

BANKING77

Method	1st run	2nd run	3rd run	Average	St. Dev.	DESCR Average
No CoT	69.1	66.23	66.23	67.18	1.66	73.15
Deep Breath	62.33	63.11	63.89	63.11	0.78	73.31
Let’s Think	66.75	67.01	64.93	66.23	1.13	71.33
Show your Thoughts	63.89	63.37	66.75	64.67	1.82	72.72
1-shot	64.15	64.93	64.41	64.49	0.40	71.32
2-shot	68.05	68.57	70.64	69.08	1.37	71.59
3-shot	67.27	69.87	68.05	68.39	1.33	73.67
BERT	94.00					

Tab. 7.6: Accuracy of BANKING77 without the use of intent descriptions, *No CoT* included (highlighted).

Similar observations can be made about our prompting technique where both the Chain of Thought and the intent descriptions are absent for the BANKING77 dataset. On this particular dataset, this method is also the best among the 0-shot methods that do not feature the intent descriptions, but not by a margin as large as it was on the CLINC-150 dataset. To be more specific, the *No CoT* method achieved a 67.18% average accuracy, a performance 0.95 percentage points better than the best prompting technique without descriptions, *Let’s*

Think, 2.61 percentage points better than *Show your Thoughts* and 4.17 percentage points better than *Deep Breath*, while maintaining a 1.66 standard deviation. As on the previous dataset, the observations made by Wei et al. (2022), Kojima et al. (2022), Wang et al. (2023), Wang et al. (2023) and others about the worse performance of Chain of Thought Prompting when compared with normal prompting on small LLMs seem to be confirmed. Now, when we compare *No CoT* with the few-shot methods without descriptions we used on this dataset we observe that the performance of *No CoT* is 2.69 percentage points better than the average accuracy of 1-shot, but still 1.9 percentage points below 2-shot, which was the best few-shot prompting technique without the descriptions added, and 1.21 percentage points below 3-shot. Also, *No CoT* was 5.97 percentage points below its description counterpart. Once more, we get an idea of the strength of descriptions when they are added to the prompt as we continuously observe that with the descriptions the results are always better.

Now for the individual runs, the best one was the first one with a 69.1% accuracy, while the second and third ones performed the same with a 66.23% each. This brings the margin between the best and worst runs to a 2.87 percentage points which is pretty average for this dataset. This shows that this technique is pretty stable between runs despite the complexity and difficulties that come with this dataset.

7.2 Error Analysis

After examining the outcomes and calculating the accuracy of our various prompting methods, we found that the best prompting technique for both datasets is the 3-shot method where we inserted 3 exemplars for in-context learning, each complete with the full prompt we discussed in earlier chapters and a response, generated by our Llama, which passes our criteria of correctly predicting the intent class of the text utterance and including a correct and reasonable Chain of Thought. Building on this foundation, we are now poised to undertake an error analysis for these two methodologies. This analysis involved examining instances of misclassification by both methods to scrutinize the Chain of Thought employed in these instances. The objective is to discern whether the inaccuracies stem from an incorrect Chain of Thought or a correct Chain of Thought relative to the input text utterance where the predicted intent class was wrong. The way that this process was done was by selecting a random sample of the misclassified responses from all three runs. This way we check the nature of the mistakes in all three runs.

The samples we utilized for this error analysis can be found in Appendix D for CLINC-150 and Appendix E for BANKING77.

7.2.1 CLINC-150

In this random sample, we found something interesting. In 60% of the errors the model made an accurate Chain of Thought based on the text, which was well worded and without any syntactical errors, but it could not make the final step in the reasoning process to correctly classify the given text utterance and instead assigning a completely wrong intent class to it. This makes us hopeful as this is something that could be mitigated with various techniques like more in-shot exemplars, better intent descriptions, or maybe even a different prompting technique. Let us not present one such case of erroneous classification.

text: do i have money in my wells fargo account for nike's

intent: balance

response: based on the text provided, the user's intent is to inquire about the current balance of their wells fargo account to determine if they have enough funds to purchase nike's products, which aligns with the "bill_balance" intent from the provided set of intents.

predicted intent : bill_balance

Error Type: Correct Chain of Thought, Wrong Classification

Now, 23.3% of the errors featured a completely wrong Chain of Thought relative to the input text utterance. Out of these responses, most of the time only a word was enough to confuse the model as to what the actual text utterance is about and it was therefore not able to form a correct Chain of Thought. The fact that this was a relatively low percentage of errors is a good thing because this shows that with the room for improvement we gained from the previous errors, the actual errors remaining would become insignificant relative to the correctly classified texts. One example of such errors is the following:

text: you need to shut it

intent: cancel

response: based on the text provided, the user's intent is to express a negative response or refusal, which aligns with the "no" intent from the previous set of intents.

predicted intent : no

Error Type: Wrong Chain of Thought

During our analysis of these errors, we came upon a third type of error, a type that we did not expect, which had nothing to do with our model and prompting techniques, as some of

the text utterances, and more specifically 16.7% of the texts in the sample, had a true intent that is close in meaning with the predicted intents, so close in fact that these text utterances could very well be multilabel. Therefore, while having a correct Chain of Thought, and having predicted a technically wrong intent class, we cannot throw any fault to our model as even humans could make these same mistakes. One such error is this:

text: is there a status update on my vacation approval

intent: pto_request_status

response: based on the text provided, the user's intent is to seek information about the current status of their vacation approval, specifically requesting an update on its status. this aligns with the "application_status" intent from the given set of intents.

predicted intent : application_status

Error Type: Correct Chain of Thought, Wrong but Plausible Classification

We have to say that we noticed a general drop in the eloquence and complexity of the Chains of Thought of these text utterances that were misclassified when compared with the corresponding eloquence and complexity of the Chains of Thought of correctly classified text utterances. This could be because the model cannot properly analyze the input text utterance and therefore cannot form a complex Chain of Thought, which could be an important factor in the misclassification of these input texts.

7.2.2 BANKING77

In this analysis, we observe a notable shift in the nature of errors compared to previous findings with the CLINC-150 dataset. Specifically, errors where the Chain of Thought was correct but the predicted intent class was incorrect accounted for only 36.7% of the total errors in the sample, a significant reduction from the 60% observed on CLINC-150. Conversely, errors involving an incorrect Chain of Thought rose to 40% of the total errors, nearly doubling the rate seen on CLINC-150. These shifts come to confirm once more both the creators of the papers and the observations made in previous chapters, that the current dataset presents more complex challenges for our model, with text and intent classifications that are more difficult to accurately interpret and understand, despite providing detailed descriptions of the intent classes. Below is an example of such errors.

The proportion of errors that could potentially be made by humans has significantly decreased to just 3.3%. This reduction is attributed to the identification of two novel error types exclusive to this dataset. The first type, accounting for 16.7% of the total errors, involves texts that

text: The rate applied to my foreign purchase was incorrect

intent: card_payment_wrong_exchange_rate

response: based on the text "the rate applied to my foreign purchase was incorrect," the user is seeking assistance with an issue related to an incorrect exchange rate being applied to a foreign purchase. therefore, the intent of the user is "exchange_rate."

predicted intent : exchange_rate

Error Type: Correct Chain of Thought, Wrong Classification

text: Where can I receive my card?

intent: order_physical_card

response: the user is asking about receiving their card, which falls under the category of getting a spare card. therefore, the intent of the user is "getting_spare_card".

predicted intent : getting_spare_card

Error Type: Wrong Chain of Thought

were incorrectly categorized from the outset, where the supposed "correct" intent class is deemed incorrect because it's unclear why these intents would be considered appropriate even from a human perspective. The second type, making up 3.3% of the total errors, includes text utterances that are not accurately labeled, although it wouldn't be accurate to say the intended correct class is wrong. An example of this is the utterance "My card is just not working at this time," which is assigned to the intent virtual_card_not_working, despite no explicit mention of the card being virtual. This was instead classified under card_not_working intent, which appears to be a more suitable categorization. Let us now present an example of such errors.

The decline in the quality of the Chain of Thought was also evident in this dataset, with numerous instances where the Chain of Thought was overly simplistic, barely meeting the criteria for being considered as such. This decline can be attributed to the increased complexity of this dataset compared to CLINC-150, as the texts in this dataset are more representative of real-life scenarios.

text: I'm still waiting on my card to be delivered.

intent: card_arrival

response: based on the text "i'm still waiting on my card to be delivered," the user is seeking information or assistance regarding the status of their card delivery. therefore, the intent of the user is "card_delivery_estimate."

predicted intent : card_delivery_estimate

Error Type: Correct Chain of Thought, Wrong but Plausible Classification

text: I live in the US but want to get a card

intent: country_support

response: based on the text "i live in the us but want to get a card," the user is seeking information about obtaining a card, which falls under the category of "supported_cards_and_currencies." therefore, the intent of the user is "supported_cards_and_currencies."

predicted intent : supported_cards_and_currencies

Error Type: Wrong label

text: My card is just not working at this time.

intent: virtual_card_not_working

response: based on the text "my card is just not working at this time," the user is experiencing issues with their card not functioning as expected, which aligns with the intent "card_not_working."

predicted intent : card_not_working

Error Type: Controversial Initial Intent

Conclusions & Future Work

8.1 Conclusions

In this thesis, we undertook a very challenging intent classification task using Large Language Models and Chain of Thought prompting. This was done to attain some explainability behind the reasons why the model classified each text to the intent class it did. The chosen model for this classification task was the Llama2 model, and more specifically the 13B variation which holds 13 billion parameters. For the purposes of this task, two open source datasets were also utilized, CLINC-150, which features 150 different intent classes across 10 domains and is relatively easy as a BERT classifier attains 96% accuracy on the test subset, and BANKING77, which features 77 intent classes, all from the banking domain, which is harder to get great results on, it was made to closely resemble real life data. Due to the extremely high inference times of the Llama model, even when quantized with 4-bit quantization, the ready-to-use test subsets of these two datasets were extremely large to be utilized in this thesis. Therefore, two new test subsets were created, both balanced with just 5 texts per intent class. This brought down inference times by a very significant margin.

In the search for the best possible results, two 0-shot prompting techniques were tested, *Let's Think* and *Deep Breath*, along with a novel prompting technique, *Show your Thoughts*. These prompts included the top-5 intent classes, as they were extracted from a BERT classifier, along with their descriptions. Each prompting technique was run a total of three times to get a better image of the stability and robustness of these three methods. The results were above expectations for the CLINC-150 dataset, where *Show your Thoughts* attained an 86% accuracy and surpassed the other two techniques. As for BANKING77 *Deep Breath* was the best among the three and obtained only a 73% accuracy. The most important aspect though is the fact that both of these techniques provided logical and eloquent Chains of Thought which can now explain the reasoning behind the model's classification of each text to its corresponding predicted intent class.

Since the 0-shot technique did not produce as satisfactory results as it would be preferable, a series of few-shot techniques were employed as providing some already answered exemplars in the context of the prompt can raise the performance of the model. This technique was inspired by the paper that introduced Chain of Thought prompting, Wei et al. (2022), where the authors hand-crafted the exemplars that were added to the context of the prompt. One

severe drawback of this technique is that there is a great variance in the performance of the model which is related to the writer of the in-shot exemplars. For this reason, in this thesis, the in-shot exemplars were extracted using the Llama model and texts that come from the training subset of each dataset.

After the extraction of the in-shot exemplars, three few-shot techniques were tested on each dataset, 1-shot, 2-shot, and 3-shot. For each dataset, the prompting technique that was utilized in the few-shot setting was the one that gave the best results in the 0-shot setting, i.e. *Show your Thoughts* for CLINC-150 and *Deep Breath* for BANKING77. The results were very promising as in both datasets the 3-shot technique became the best performing technique, either 0-shot or few-shot, surpassing all other prompting techniques. In CLINC-150 3-shot achieved a remarkable 88.88% while in BANKING it achieved 73.67%, which may not seem like a lot, but it is a promising result as it comes closer to the 74% accuracy that Loukas et al. (2023) attained using GPT-4 in a classification task using the same dataset.

The Chains of Thought that were created using the 3-shot technique were more eloquent and complex on CLINC-150, where a correct reasoning path was chosen almost all of the time. On BANKING77, Llama faces challenges in generating complex Chains of Thought for texts in this dataset. These hurdles extend to the eloquence of the selected Chains of Thought. One factor contributing to these difficulties could be the nature of the texts themselves. They are designed to be more challenging and resemble real-life data, which may present obstacles to Llama's performance. Additionally, the in-shot exemplars provided may not be sufficient for Llama to achieve better results and a wider margin from the 0-shot results. Moreover, the number of in-shot exemplars, which stands at three, might be too limited for this particular dataset, further hindering Llama's performance.

8.2 Future Work

Chain of Thought Prompting is a relatively new technique as the paper that introduced it only came out in 2022. Intent classification with Chain of Thought Prompting is a novel problem as, according to the available information, it has not been explored before this thesis. Therefore, the areas of experimentation that could be explored are practically limitless.

The first and easiest way that this series of experiments could expand is by utilizing a different model for inference, maybe a bigger one, like Llama2-70B, or Mistral 8x7b, both of which show

superior performance to the Llama2-13B across almost every benchmark. Another expansion method could be calculating the accuracy on the whole test subset of each dataset and thus calculating the definitive accuracy on each dataset. A different path that could be explored is to run all the experiments that were previously analyzed for more runs so that the average accuracy would be more representative when compared with the 3 runs that were used in this thesis.

One way that this experimentation could continue is by further exploring few-shot prompting methods. This could happen in a few different ways. One possible way is by continuing the work we described above and finding more exemplars to be utilized in a 4-shot, 5-shot, or more scenario, where the exemplars continue to fit the criteria that were set in previous chapters. A different approach would be to continue the series of experiments that were analyzed previously by introducing each time the in-shot exemplars randomly from a pool of exemplars that were chosen so that they fit the criteria that were set previously.

A completely different path that could be explored is the utilization of more complex techniques that incite Chain of Thought prompting. This could be done by testing different Chain-of-Thought-inciting phrases, like Plan-and-Solve prompting, a technique that Wang et al. (2023) found to work better than *Let's Think* in a lot of reasoning problems. A different approach is Plan and Solve prompting by Wang et al. (2023), which the authors of the paper describe as a *self-ensembling* method since it samples k different reasoning paths and the most frequent one is taken as the answer. This technique could prove superior as it can be used with all of the 0-shot prompting techniques that were explored in this thesis and it is considered maybe the most promising technique with the possibility to yield the highest improvement in performance.

Bibliography

- A. M. P. Aggarwal and M. Y. Yang (2023). “Let’s Sample Step by Step: Adaptive-Consistency for Efficient Reasoning and Coding with LLMs”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Ed. by H. Bouamor, J. Pino, and K. Bali. Association for Computational Linguistics, pp. 12375–12396. URL: <https://aclanthology.org/2023.emnlp-main.761>.
- J. Ainslie, J. Lee-Thorp, M. de Jong, et al. (2023). “GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, pp. 4895–4901. URL: <https://aclanthology.org/2023.emnlp-main.298>.
- A. G. Bouzaki (2023). *Enhancing Intent Classification via Zero-shot and Few-shot ChatGPT Prompting Engineering: Generating training data or directly detecting intents?* URL: <https://pergamos.lib.uoa.gr/uoa/dl/object/3359481/file.pdf>.
- T. Brown, B. Mann, N. Ryder, et al. (2020). “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
- I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić (July 2020). “Efficient Intent Detection with Dual Sentence Encoders”. In: *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Ed. by T.-H. Wen, A. Celikyilmaz, Z. Yu, et al. Online: Association for Computational Linguistics, pp. 38–45. URL: <https://aclanthology.org/2020.nlp4convai-1.5>.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>.

- B. Jacob, S. Kligys, B. Chen, et al. (2017). “Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713. URL: <https://api.semanticscholar.org/CorpusID:39867659>.
- T. Kojima, S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa (2022). “Large Language Models are Zero-Shot Reasoners”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, et al. Vol. 35. Curran Associates, Inc., pp. 22199–22213. URL: <https://openreview.net/pdf?id=e2TBb5y0yFf>.
- S. Larson, A. Mahendran, J. J. Peper, et al. (Nov. 2019). “An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Hong Kong, China: Association for Computational Linguistics, pp. 1311–1316. URL: <https://aclanthology.org/D19-1131>.
- Y. Liu, M. Ott, N. Goyal, et al. (2019). “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *CoRR* abs/1907.11692. arXiv: 1907.11692. URL: <http://arxiv.org/abs/1907.11692>.
- L. Loukas, I. Stogiannidis, O. Diamantopoulos, P. Malakasiotis, and S. Vassos (2023). “Making LLMs Worth Every Penny: Resource-Limited Text Classification in Banking”. In: *Proceedings of the Fourth ACM International Conference on AI in Finance*. ICAIF ’23. New York, NY, USA: Association for Computing Machinery, pp. 392–400. URL: <https://doi.org/10.1145/3604237.3626891>.
- L. C. Magister, J. Mallinson, J. Adamek, E. Malmi, and A. Severyn (July 2023). “Teaching Small Language Models to Reason”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 1773–1781. URL: <https://aclanthology.org/2023.acl-short.151>.
- OpenAI (2022). *Introducing ChatGPT*. URL: <https://openai.com/blog/chatgpt>.
- L. Ouyang, J. Wu, X. Jiang, et al. (2022). “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, et al. Vol. 35. Curran Associates, Inc., pp. 27730–27744. URL: https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf.

- S. Parikh, M. Tiwari, P. Tumbade, and Q. Vohra (July 2023). “Exploring Zero and Few-shot Techniques for Intent Classification”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. Ed. by S. Sitaram, B. Beigman Klebanov, and J. D. Williams. Toronto, Canada: Association for Computational Linguistics, pp. 744–751. URL: <https://aclanthology.org/2023.acl-industry.71>.
- A. Radford, J. Wu, R. Child, et al. (2019). “Language Models are Unsupervised Multitask Learners”. In: URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- C. Raffel, N. Shazeer, A. Roberts, et al. (2020). “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: <http://jmlr.org/papers/v21/20-074.html>.
- J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu (2021). “RoFormer: Enhanced Transformer with Rotary Position Embedding”. In: *CoRR* abs/2104.09864. eprint: 2104.09864. URL: <https://arxiv.org/abs/2104.09864>.
- H. Touvron, L. Martin, K. Stone, et al. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: 2307.09288 [cs.CL]. URL: <https://arxiv.org/pdf/2307.09288.pdf>.
- M. Turpin, J. Michael, E. Perez, and S. Bowman (2023). “Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Neumann, A. Globerson, et al. Vol. 36. Curran Associates, Inc., pp. 74952–74965. URL: <https://arxiv.org/pdf/2305.04388.pdf>.
- A. Vaswani, N. Shazeer, N. Parmar, et al. (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- L. Wang, W. Xu, Y. Lan, et al. (2023). “Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by A. Rogers, J. L. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, pp. 2609–2634. URL: <https://doi.org/10.18653/v1/2023.acl-long.147>.

- X. Wang, J. Wei, D. Schuurmans, et al. (2023). “Self-Consistency Improves Chain of Thought Reasoning in Language Models”. In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. URL: <https://openreview.net/pdf?id=1PL1NIMMrw>.
- J. Wei, X. Wang, D. Schuurmans, et al. (2022). “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, et al. Vol. 35. Curran Associates, Inc., pp. 24824–24837. URL: <https://arxiv.org/abs/2201.11903>.
- C. Yang, X. Wang, Y. Lu, et al. (2023). “Large Language Models as Optimizers”. In: *CoRR* abs/2309.03409. arXiv: 2309.03409. URL: <https://doi.org/10.48550/arXiv.2309.03409>.

Prompt Examples

In this section, we present 2 prompt examples per prompting technique, one with the descriptions included, and one not, for both datasets. This is done in order for the reader to fully understand the way that these prompts are set up and exactly what they look like.

CLINC-150 - Deep Breath Prompt 1: with Descriptions

We have the following set of intents along with their descriptions:

- * report_fraud: The intent `"report_fraud"` involves seeking guidance or providing information regarding suspected fraudulent activities, such as unauthorized transactions or deceptive practices, to appropriate authorities for investigation.
- * lost_luggage: The intent `"lost_luggage"` involves seeking help or information regarding misplaced or missing baggage during travel, such as reporting the loss, inquiring about its current status, or requesting assistance in locating lost luggage.
- * expiration_date: The intent `"expiration_date"` involves seeking information or clarification regarding the expiry date of a product, document, or any time-sensitive item.
- * report_lost_card: The intent `"report_lost_card"` pertains to seeking assistance or notifying relevant parties about a misplaced or lost credit/debit card and initiating necessary measures to address the situation, such as card replacement or security measures.
- * damaged_card: The intent `"damaged_card"` encompasses requests or inquiries related to obtaining assistance or information regarding a credit, debit, or identification card that has been compromised, broken, or damaged.

A user wrote the following text: `"i would like to report my american express card stolen"`. The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

CLINC-150 - Deep Breath Prompt 2: without Descriptions

We have the following set of intents: `['directions', 'accept_reservations', 'restaurant_reservation', 'book_flight', 'confirm_reservation']`.

A user wrote the following text: `"any clue as to if taco bell takes reservations"`.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

CLINC-150 - Let's Think Prompt 1: with Descriptions

We have the following set of intents along with their descriptions:

- * whisper_mode: The intent `"whisper_mode"` pertains to requests or settings related to a device or application being configured to operate in a discreet or quiet mode, either engaging said mode or disengaging it.
- * repeat: The intent `"repeat"` involves asking for the duplication or reiteration of a statement, command, or information previously provided, seeking clarification or emphasis.
- * translate: The intent `"translate"` involves requesting assistance in converting a word or phrase from one language to another, seeking clarification or understanding in a language distinct from the original.
- * change_language: The intent `"change_language"` involves seeking guidance or instructions on how to switch the language settings to a different one on a device or platform.
- * spelling: The intent `"spelling"` involves inquiries or requests related to the correct spelling of a word or phrase, typically seeking clarification on the accurate arrangement of letters in a specific sequence.

A user wrote the following text: "would you understand me if i speak to you in navajo". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's think step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

CLINC-150 - Let's Think Prompt 2: without Descriptions

We have the following set of intents: `['interest_rate', 'exchange_rate', 'transfer', 'min_payment', 'measurement_conversion']`.

A user wrote the following text: "what is the exchange rate between euros and pesos".

What was the intent of the user when they wrote this text? Let's think step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

CLINC-150 - Show your Thoughts Prompt 1: with Descriptions

We have the following set of intents along with their descriptions:

- * confirm_reservation: The intent `"confirm_reservation"` involves seeking verification or assurance regarding a previously made reservation, typically inquiring about the details or status of a booked service, such as a hotel room or restaurant table.
- * accept_reservations: The intent `"accept_reservations"` involves expressing a desire to make a reservation or inquiring about a business or service's capability to accommodate advance bookings for a specified date or time.
- * restaurant_reservation: The intent `"restaurant_reservation"` involves making inquiries or requests related to securing a table at a dining establishment for a specific date and time.
- * book_flight: The intent `"book_flight"` involves users expressing an intention to reserve or secure a seat on a flight, typically by inquiring about available flights, prices, or the booking process for air travel.
- * directions: The intent `"directions"` involves seeking guidance or information on navigating to a specific destination, whether it's asking for directions to a particular place or inquiring about the route to reach a specific location.

A user wrote the following text: "any clue as to if taco bell takes reservations". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

CLINC-150 - Show your Thoughts Prompt 2: without Descriptions

We have the following set of intents: ['freeze_account', 'pay_bill', 'account_blocked', 'balance', 'report_fraud'].

A user wrote the following text: "can you check what the problem with my account is, it seems to be blocked".

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

CLINC-150 - 1-shot Prompt 1: with Descriptions

#Question: We have the following set of intents along with their descriptions:

- * credit_limit: The intent `"credit_limit"` involves inquiries related to the maximum amount of credit available to an individual or entity, typically seeking information or clarification regarding the assigned credit limit on a financial account.
- * freeze_account: The intent `"freeze_account"` involves seeking guidance or initiating a request to suspend or lock a financial account temporarily, typically due to concerns about security or unauthorized access.
- * balance: The intent `"balance"` pertains to inquiries or requests related to checking the remaining amount or status of a financial account, be it a bank account, prepaid card, or any other financial instrument, where users seek information about their current balance.

- * credit_limit_change: The intent "credit_limit_change" relates to requests for modifications or updates to the maximum amount of credit available on an account, often involving inquiries about the process or requirements for adjusting the assigned credit limit.
- * improve_credit_score: The intent "improve_credit_score" involves seeking guidance or information on strategies, tips, or actions to enhance one's credit score or addressing inquiries related to creditworthiness improvement.

A user wrote the following text: "how do i request a credit limit increase on my visa". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Since the user is asking how to request an increase in their credit limit on their Visa account, which falls under the category of modifying or updating the assigned credit limit, the intent is "credit_limit_change".

#Question: We have the following set of intents along with their descriptions:

- * whisper_mode: The intent "whisper_mode" pertains to requests or settings related to a device or application being configured to operate in a discreet or quiet mode, either engaging said mode or disengaging it.
- * change_speed: The intent "change_speed" involves seeking information or guidance on adjusting the speed settings, such as playback speed or talking speed, on a device or application.
- * change_volume: The intent "change_volume" pertains to inquiries or commands related to adjusting the audio output levels, whether increasing or decreasing the volume.
- * repeat: The intent "repeat" involves asking for the duplication or reiteration of a statement, command, or information previously provided, seeking clarification or emphasis.
- * reminder: The intent "reminder" pertains to inquiries regarding a reminder list that a user set up in order to manage their reminders.

A user wrote the following text: "speak more quickly". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

CLINC-150 - 1-shot Prompt 2: without Descriptions

#Question: We have the following set of intents: ['rewards_balance', 'exchange_rate', 'balance', 'transfer', 'income']. A user wrote the following text: "deposit ten bucks from my chase account to my ally account". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: The intent of the user when they wrote the text "deposit ten bucks from my chase account to my ally account" is "transfer". This is because the user is instructing a financial institution to move funds from one account to another, which aligns with the "transfer" intent in the provided set.

#Question: We have the following set of intents: ['transfer', 'calculator', 'reset_settings', 'cancel', 'sync_device']. A user wrote the following text: "please remove current settings and go back to factory settings now". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

CLINC-150 - 2-shot Prompt 2: with Descriptions

#Question: We have the following set of intents along with their descriptions:

* credit_limit: The intent "credit_limit" involves inquiries related to the maximum amount of credit available to an individual or entity, typically seeking information or clarification regarding the assigned credit limit on a financial account.

* freeze_account: The intent "freeze_account" involves seeking guidance or initiating a request to suspend or lock a financial account temporarily, typically due to concerns about security or unauthorized access.

* balance: The intent "balance" pertains to inquiries or requests related to checking the remaining amount or status of a financial account, be it a bank account, prepaid card, or any other financial instrument, where users seek information about their current balance.

* credit_limit_change: The intent "credit_limit_change" relates to requests for modifications or updates to the maximum amount of credit available on an account, often involving inquiries about the process or requirements for adjusting the assigned credit limit.

* improve_credit_score: The intent "improve_credit_score" involves seeking guidance or information on strategies, tips, or actions to enhance one's credit score or addressing inquiries related to creditworthiness improvement.

A user wrote the following text: "how do i request a credit limit increase on my visa". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Since the user is asking how to request an increase in their credit limit on their Visa account, which falls under the category of modifying or updating the assigned credit limit, the intent is ""credit_limit_change"".

#Question: We have the following set of intents along with their descriptions:

- * how_busy: The intent "how_busy" relates to queries or requests for information regarding the current level of activity or occupancy, often inquiring about the busyness of a place, event, or specific timeframe.
- * food_last: The intent "food_last" involves seeking information about the remaining time until a food or beverage goes bad and cannot be consumed any more.
- * traffic: The intent "traffic" involves seeking information or assistance related to current traffic conditions, such as updates on congestion, road closures, or alternative routes for navigation.
- * meeting_schedule: The intent "meeting_schedule" pertains to seeking information about the time of a particular meeting or the meetings scheduled for a particular day.
- * distance: The intent "distance" pertains to inquiries about the measurement or estimation of the distance between two points, locations, or objects, seeking information on either the spatial separation or time separation in a given context.

A user wrote the following text: ""how long will the wait be at noon if we went to nagomi"". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text provided, the user's intent is to seek information about the current level of activity or occupancy of Nagomi, as they inquire about the expected wait time at noon. Therefore, the intent of the user "how_busy".

#Question: We have the following set of intents along with their descriptions:

- * change_volume: The intent "change_volume" pertains to inquiries or commands related to adjusting the audio output levels, whether increasing or decreasing the volume.
- * reminder: The intent "reminder" pertains to inquiries regarding a reminder list that a user set up in order to manage their reminders.
- * change_speed: The intent "change_speed" involves seeking information or guidance on adjusting the speed settings, such as playback speed or talking speed, on a device or application.
- * repeat: The intent "repeat" involves asking for the duplication or reiteration of a statement, command, or information previously provided, seeking clarification or emphasis.
- * whisper_mode: The intent "whisper_mode" pertains to requests or settings related to a device or application being configured to operate in a discreet or quiet mode, either engaging said mode or disengaging it.

A user wrote the following text: "go faster". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

CLINC-150 - 2-shot Prompt 2: without Descriptions

#Question: We have the following set of intents: ['rewards_balance', 'exchange_rate', 'balance', 'transfer', 'income']. A user wrote the following text: "deposit ten bucks from my chase account to my ally account". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: The intent of the user when they wrote the text "deposit ten bucks from my chase account to my ally account" is "transfer". This is because the user is instructing a financial institution to move funds from one account to another, which aligns with the "transfer" intent in the provided set.

#Question: We have the following set of intents: ['transactions', 'order_status', 'change_accent', 'share_location', 'order_checks']. A user wrote the following text: "where was my package last at". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: The intent of the user when they wrote the text "where was my package last at" is likely "order_status". This is because the user is asking about the location of their package, which is a common concern when tracking the status of an order. The wording of the question suggests that the user is looking for information about the current location of their package, which aligns with the "order_status" intent.

#Question: We have the following set of intents: ['taxes', 'rollover_401k', 'make_call', 'insurance_change', 'insurance']. A user wrote the following text: "is there a way to alter my insurance policy". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

CLINC-150 - 3-shot Prompt 1: with Descriptions

#Question: We have the following set of intents along with their descriptions:

- * credit_limit: The intent "credit_limit" involves inquiries related to the maximum amount of credit available to an individual or entity, typically seeking information or clarification regarding the assigned credit limit on a financial account.
- * freeze_account: The intent "freeze_account" involves seeking guidance or initiating a request to suspend or lock a financial account temporarily, typically due to concerns about security or unauthorized access.
- * balance: The intent "balance" pertains to inquiries or requests related to checking the remaining amount or status of a financial account, be it a bank account, prepaid card, or any other financial instrument, where users seek information about their current balance.
- * credit_limit_change: The intent "credit_limit_change" relates to requests for modifications or updates to the maximum amount of credit available on an account, often involving inquiries about the process or requirements for adjusting the assigned credit limit.
- * improve_credit_score: The intent "improve_credit_score" involves seeking guidance or information on strategies, tips, or actions to enhance one's credit score or addressing inquiries related to creditworthiness improvement.

A user wrote the following text: "how do i request a credit limit increase on my visa". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Since the user is asking how to request an increase in their credit limit on their Visa account, which falls under the category of modifying or updating the assigned credit limit, the intent is "credit_limit_change".

#Question: We have the following set of intents along with their descriptions:

- * how_busy: The intent "how_busy" relates to queries or requests for information regarding the current level of activity or occupancy, often inquiring about the busyness of a place, event, or specific timeframe.
- * food_last: The intent "food_last" involves seeking information about the remaining time until a food or beverage goes bad and cannot be consumed any more.
- * traffic: The intent "traffic" involves seeking information or assistance related to current traffic conditions, such as updates on congestion, road closures, or alternative routes for navigation.
- * meeting_schedule: The intent "meeting_schedule" pertains to seeking information about the time of a particular meeting or the meetings scheduled for a particular day.
- * distance: The intent "distance" pertains to inquiries about the measurement or estimation of the distance between two points, locations, or objects, seeking information on either the spatial separation or time separation in a given context.

A user wrote the following text: "how long will the wait be at noon if we went to nagomi". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text provided, the user's intent is to seek information about the current level of activity or occupancy of Nagomi, as they inquire about the expected wait time at noon. Therefore, the intent of the user "how_busy".

#Question: We have the following set of intents along with their descriptions:

- * routing: The intent "routing" pertains to requesting assistance with a specific bank's routing number.

- * what_can_i_ask_you: The intent "what_can_i_ask_you" pertains to seeking guidance or suggestions on the types of questions or topics that can be posed.

- * what_are_your_hobbies: The intent "what_are_your_hobbies" involves prompting someone to share information about their personal interests and activities, offering insights into the individual's leisure pursuits and pastimes.

- * how_old_are_you: The intent "how_old_are_you" involves inquiries or comments seeking information about the age of a person, entity, or object.

- * travel_suggestion: The intent "travel_suggestion" involves seeking recommendations or advice related to travel plans, such as destination suggestions, activity recommendations, or guidance on creating an enjoyable itinerary for an upcoming trip.

A user wrote the following text: "please tell me your hobbies". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the given text, the intent of the user when they wrote "please tell me your hobbies" is to ask for the other party's personal interests and activities, which aligns with the "what_are_your_hobbies" intent from the provided set of intents.

#Question: We have the following set of intents along with their descriptions:

- * gas_type: The intent "gas_type" involves seeking information about the specific type or grade of fuel required for a vehicle or a related inquiry about available fuel options.

- * insurance: The intent "insurance" pertains to inquiries or requests related to obtaining information about insurance coverage or the specifics of the insurance plan of the user.

- * change_speed: The intent "change_speed" involves seeking information or guidance on adjusting the speed settings, such as playback speed or talking speed, on a device or application.

- * insurance_change: The intent "insurance_change" involves seeking guidance or information related to making alterations or updates to an existing insurance policy, such as modifying coverage details or changing insurance plans.

- * vaccines: The intent "vaccines" involves inquiries about the necessary vaccination in order to travel to a country.

A user wrote the following text: "do i need a vaccine for the congo". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

CLINC-150 - 3-shot Prompt 4: without Descriptions

#Question: We have the following set of intents: ['rewards_balance', 'exchange_rate', 'balance', 'transfer', 'income']. A user wrote the following text: "deposit ten bucks from my chase account to my ally account". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: The intent of the user when they wrote the text "deposit ten bucks from my chase account to my ally account" is "transfer". This is because the user is instructing a financial institution to move funds from one account to another, which aligns with the "transfer" intent in the provided set.

#Question: We have the following set of intents: ['transactions', 'order_status', 'change_accent', 'share_location', 'order_checks']. A user wrote the following text: "where was my package last at". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: The intent of the user when they wrote the text "where was my package last at" is likely "order_status". This is because the user is asking about the location of their package, which is a common concern when tracking the status of an order. The wording of the question suggests that the user is looking for information about the current location of their package, which aligns with the "order_status" intent.

#Question: We have the following set of intents: ['change_user_name', 'restaurant_reservation', 'accept_reservations', 'confirm_reservation', 'book_flight']. A user wrote the following text: "do they take reservations at red robin". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: The intent of the user when they wrote "do they take reservations at red robin" is likely "accept_reservations". This is because the user is asking if 'red robin' takes reservations, which aligns with the "accept_reservations" intent.

#Question: We have the following set of intents: ['smart_home', 'schedule_meeting', 'meeting_schedule', 'restaurant_reservation', 'book_hotel']. A user wrote the following text: "friday 5pm schedule a room for meeting". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

BANKING77 - Deep Breath Prompt 1: with Descriptions

We have the following set of intents along with their descriptions:

- * `reverted_card_payment?`: The intent `"reverted_card_payment?"` pertains to inquiries or concerns regarding transactions that have been canceled or reversed on a card, often seeking clarification on the reasons, processes, or implications of such reversals.
- * `transaction_charged_twice`: The intent `"transaction_charged_twice"` corresponds to inquiries or concerns regarding a situation where a user's account shows that a single transaction has been processed and debited more than once, leading to confusion and a need for clarification or resolution.
- * `request_refund`: The intent `"request_refund"` refers to an inquiry or action taken by a user to seek the return of funds for a transaction or service that did not meet their expectations or requirements.
- * `pending_card_payment`: The intent `"pending_card_payment"` refers to inquiries or concerns regarding card payments that have been initiated but have not yet been completed or cleared, often seeking information about the status or expected completion time of the transaction.
- * `Refund_not_showing_up`: The intent `"Refund_not_showing_up"` pertains to inquiries or concerns regarding the visibility or status of a processed refund that has not yet appeared in the individual's account or financial statements.

A user wrote the following text: "My transaction to pay for an item was returned to my account.". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

BANKING77 - Deep Breath Prompt 2: without Descriptions

We have the following set of intents: `['supported_cards_and_currencies', 'card_delivery_estimate', 'card_payment_not_recognised', 'card_acceptance', 'country_support']`.

A user wrote the following text: "Can the card be mailed and used in Europe?".

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

BANKING77 - Let's Think Prompt 1: with Descriptions

We have the following set of intents along with their descriptions:

- * pin_blocked: The intent "pin_blocked" pertains to inquiries or concerns regarding the inability to access or use a card due to the personal identification number (PIN) being locked or disabled.
- * beneficiary_not_allowed: The intent "beneficiary_not_allowed" pertains to inquiries or statements regarding the prohibition or restriction of adding or designating a certain individual or entity as a recipient of benefits or funds.
- * direct_debit_payment_not_recognised: The intent "direct_debit_payment_not_recognised" refers to inquiries or concerns regarding a direct debit transaction that the individual does not acknowledge or understand, often implying a need for clarification or resolution.
- * card_arrival: The intent "card_arrival" refers to a request for assistance in tracking a card or a question about its current location or time of arrival.
- * passcode_forgotten: The intent "passcode_forgotten" refers to inquiries or concerns expressed when an individual is unable to recall or access their personal identification number or security code, seeking assistance or solutions to regain entry into a secured account or device.

A user wrote the following text: "I need a new passcode.". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's think step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

BANKING77 - Let's Think Prompt 2: without Descriptions

We have the following set of intents: ['reverted_card_payment?', 'transaction_charged_twice', 'request_refund', 'pending_card_payment', 'Refund_not_showing_up'].

A user wrote the following text: "My transaction to pay for an item was returned to my account.".

What was the intent of the user when they wrote this text? Let's think step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

BANKING77 - Show your Thoughts Prompt 1: with Descriptions

We have the following set of intents along with their descriptions:

- * why_verify_identity: The intent `"why_verify_identity"` refers to inquiries seeking clarification or understanding regarding the reasons behind the need for identity verification procedures.
- * verify_my_identity: The intent `"verify_my_identity"` pertains to inquiries or requests for the procedures and necessary steps to confirm one's personal identification, often as a security measure or prerequisite for accessing services or information.
- * order_physical_card: The intent `"order_physical_card"` pertains to queries or actions related to requesting a new physical card, typically involving inquiries about the process, necessary steps, or requirements for obtaining a tangible debit or credit card from a financial institution or service provider.
- * unable_to_verify_identity: The intent `"unable_to_verify_identity"` corresponds to expressions of concern or inquiries related to difficulties in confirming personal identification during a verification process.
- * card_acceptance: The intent `"card_acceptance"` pertains to inquiries or clarifications regarding the acceptance or recognition of a card at various establishments or through different transaction methods.

A user wrote the following text: `"What do I need to verify my identity"`. The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

BANKING77 - Show your Thoughts Prompt 2: without Descriptions

We have the following set of intents: `['wrong_amount_of_cash_received', 'exchange_via_app', 'supported_cards_and_currencies', 'fiat_currency_support', 'exchange_charge']`.

A user wrote the following text: `"Tell me how I can exchange between different currency."`.

What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

BANKING77 - 1-shot Prompt 1: with Descriptions

#Question: We have the following set of intents along with their descriptions:

- * `unable_to_verify_identity`: The intent `"unable_to_verify_identity"` corresponds to expressions of concern or inquiries related to difficulties in confirming personal identification during a verification process.
- * `card_payment_not_recognised`: The intent `"card_payment_not_recognised"` pertains to inquiries or concerns about a payment made with a card that is not acknowledged or appearing as valid in transaction records.
- * `verify_my_identity`: The intent `"verify_my_identity"` pertains to inquiries or requests for the procedures and necessary steps to confirm one's personal identification, often as a security measure or prerequisite for accessing services or information.
- * `edit_personal_details`: The intent `"edit_personal_details"` signifies a user's desire to modify or update their own personal information within a system or service.
- * `why_verify_identity`: The intent `"why_verify_identity"` refers to inquiries seeking clarification or understanding regarding the reasons behind the need for identity verification procedures.

A user wrote the following text: "Why is there a verification to identity". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text `"Why is there a verification to identity"`, the user is seeking clarification and understanding about the reasons behind the need for identity verification procedures. Therefore the intent of the user is `"why_verify_identity"`.

#Question: We have the following set of intents along with their descriptions:

- * `pending_transfer`: The intent `"pending_transfer"` refers to inquiries or concerns about the status of a financial transaction that has been initiated but not yet completed, often seeking clarification on the expected completion time or reasons for the delay.
- * `beneficiary_not_allowed`: The intent `"beneficiary_not_allowed"` pertains to inquiries or statements regarding the prohibition or restriction of adding or designating a certain individual or entity as a recipient of benefits or funds.
- * `balance_not_updated_after_bank_transfer`: The intent `"balance_not_updated_after_bank_transfer"` refers to inquiries or concerns regarding the non-updating or delayed updating of an account balance following a bank transfer transaction.
- * `transfer_timing`: The intent `"transfer_timing"` pertains to inquiries or concerns regarding the specific timing, duration, or schedule associated with transferring funds between accounts or entities.
- * `transfer_not_received_by_recipient`: The intent `"transfer_not_received_by_recipient"` pertains to inquiries or concerns about a financial transfer that the intended recipient has not acknowledged receiving, focusing on understanding the reasons for the delay or identifying the whereabouts of the funds.

A user wrote the following text: "Why doesn't my balance reflect my transfer". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

BANKING77 - 1-shot Prompt 2: without Descriptions

#Question: We have the following set of intents: ['extra_charge_on_statement', 'reverted_card_payment?', 'transaction_charged_twice', 'why_verify_identity', 'card_payment_not_recognised']. A user wrote the following text: "There are two of the same charge on my statement.". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: The intent of the user when writing the text "There are two of the same charge on my statement" is to report an issue with a duplicate transaction, specifically the intent "transaction_charged_twice".

#Question: We have the following set of intents: ['supported_cards_and_currencies', 'verify_source_of_funds', 'wrong_amount_of_cash_received', 'fiat_currency_support', 'receiving_money']. A user wrote the following text: "What is the source of my funds.". What was the intent of the user when they wrote this text? Show your thoughts, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

BANKING77 - 2-shot Prompt 1: with Descriptions

#Question: We have the following set of intents along with their descriptions:

* unable_to_verify_identity: The intent "unable_to_verify_identity" corresponds to expressions of concern or inquiries related to difficulties in confirming personal identification during a verification process.

* card_payment_not_recognised: The intent "card_payment_not_recognised" pertains to inquiries or concerns about a payment made with a card that is not acknowledged or appearing as valid in transaction records.

* verify_my_identity: The intent "verify_my_identity" pertains to inquiries or requests for the procedures and necessary steps to confirm one's personal identification, often as a security measure or prerequisite for accessing services or information.

* edit_personal_details: The intent "edit_personal_details" signifies a user's desire to modify or update their own personal information within a system or service.

* why_verify_identity: The intent "why_verify_identity" refers to inquiries seeking clarification or understanding regarding the reasons behind the need for identity verification procedures.

A user wrote the following text: "Why is there a verification to identity". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text "Why is there a verification to identity", the user is seeking clarification and understanding about the reasons behind the need for identity verification procedures. Therefore the intent of the user is "why_verify_identity".

#Question: We have the following set of intents along with their descriptions:

- * request_refund: The intent "request_refund" refers to an inquiry or action taken by a user to seek the return of funds for a transaction or service that did not meet their expectations or requirements.
- * cancel_transfer: The intent "cancel_transfer" signifies the desire to halt or reverse a previously initiated monetary transaction, often motivated by a change in circumstances or an error in the original instruction.
- * edit_personal_details: The intent "edit_personal_details" signifies a user's desire to modify or update their own personal information within a system or service.
- * failed_transfer: The intent "failed_transfer" refers to inquiries or complaints regarding unsuccessful attempts to move funds from one account or entity to another.
- * transfer_not_received_by_recipient: The intent "transfer_not_received_by_recipient" pertains to inquiries or concerns about a financial transfer that the intended recipient has not acknowledged receiving, focusing on understanding the reasons for the delay or identifying the whereabouts of the funds.

A user wrote the following text: "I would like help to revert a transaction I did this morning". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text "I would like help to revert a transaction I did this morning", the user is seeking assistance to reverse a recent transaction. Therefore the intent of the user is "cancel_transfer".

#Question: We have the following set of intents along with their descriptions:

- * atm_support: The intent "atm_support" pertains to seeking help or information related to the functioning, issues, or usage of an ATM.
- * declined_cash_withdrawal: The intent "declined_cash_withdrawal" signifies inquiries or concerns related to an unsuccessful attempt to withdraw cash from an account, typically seeking clarification or resolution for the transaction being denied.

- * `wrong_amount_of_cash_received`: The intent `"wrong_amount_of_cash_received"` refers to expressing concerns or seeking clarification when the amount of money received during a transaction does not match the expected value.
- * `virtual_card_not_working`: The intent `"virtual_card_not_working"` encompasses inquiries or reports of issues related to the functionality or acceptance of a virtual card, typically focusing on situations where the card is not being processed or accepted as expected during online or in-app transactions.
- * `card_swallowed`: The intent `"card_swallowed"` pertains to inquiries or concerns related to a bank card being unexpectedly retained by an ATM or payment device.

A user wrote the following text: `"Where are the locations of ATMs that accept this card?"`. The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

BANKING77 - 2-shot Prompt 2: without Descriptions

#Question: We have the following set of intents: `['cancel_transfer', 'failed_transfer', 'transfer_not_received_by_recipient', 'edit_personal_details', 'request_refund']`.

A user wrote the following text: `"Help! I need to cancel a transaction."`.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text `"Help! I need to cancel a transaction"`, the intent of the user is to request a cancellation of a transfer, specifically a transfer that has already been initiated and needs to be stopped or reversed, which aligns with the intent `"cancel_transfer"` from the provided set.

#Question: We have the following set of intents: `['extra_charge_on_statement', 'card_swallowed', 'cash_withdrawal_charge', 'card_payment_fee_charged', 'declined_cash_withdrawal']`.

A user wrote the following text: `"What is your policy on withdrawing cash from an ATM? There is a fee on my account that wasn't there before I paid for my groceries."`.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text, the intent of the user is to ask about the policy on withdrawing cash from an ATM, specifically asking about any fees associated with such a transaction, which aligns with the `"cash_withdrawal_charge"` intent.

#Question: We have the following set of intents: ['automatic_top_up', 'top_up_limits', 'top_up_failed', 'pending_top_up', 'exchange_charge'].

A user wrote the following text: "Is there a limit to top-ups?"

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

BANKING77 - 3-shot Prompt 1: with Descriptions

#Question: We have the following set of intents along with their descriptions:

- * `unable_to_verify_identity`: The intent `"unable_to_verify_identity"` corresponds to expressions of concern or inquiries related to difficulties in confirming personal identification during a verification process.
- * `card_payment_not_recognised`: The intent `"card_payment_not_recognised"` pertains to inquiries or concerns about a payment made with a card that is not acknowledged or appearing as valid in transaction records.
- * `verify_my_identity`: The intent `"verify_my_identity"` pertains to inquiries or requests for the procedures and necessary steps to confirm one's personal identification, often as a security measure or prerequisite for accessing services or information.
- * `edit_personal_details`: The intent `"edit_personal_details"` signifies a user's desire to modify or update their own personal information within a system or service.
- * `why_verify_identity`: The intent `"why_verify_identity"` refers to inquiries seeking clarification or understanding regarding the reasons behind the need for identity verification procedures.

A user wrote the following text: "Why is there a verification to identity". The intent of this text is definitely one from the five intents in the previous set.

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text `"Why is there a verification to identity"`, the user is seeking clarification and understanding about the reasons behind the need for identity verification procedures. Therefore the intent of the user is `"why_verify_identity"`.

#Question: We have the following set of intents along with their descriptions:

- * `request_refund`: The intent `"request_refund"` refers to an inquiry or action taken by a user to seek the return of funds for a transaction or service that did not meet their expectations or requirements.
- * `cancel_transfer`: The intent `"cancel_transfer"` signifies the desire to halt or reverse a previously initiated monetary transaction, often motivated by a change in circumstances or an error in the original instruction.
- * `edit_personal_details`: The intent `"edit_personal_details"` signifies a user's desire to modify or update their own personal information within a system or service.
- * `failed_transfer`: The intent `"failed_transfer"` refers to inquiries or complaints regarding unsuccessful attempts to move funds from one account or entity to another.
- * `transfer_not_received_by_recipient`: The intent `"transfer_not_received_by_recipient"` pertains to inquiries or concerns about a financial transfer that the intended recipient has not acknowledged receiving, focusing on understanding the reasons for the delay or identifying the whereabouts of the funds.

A user wrote the following text: "I would like help to revert a transaction I did this morning". The intent of this text is definitely one from the five intents in the previous set. What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text "I would like help to revert a transaction I did this morning", the user is seeking assistance to reverse a recent transaction. Therefore the intent of the user is "cancel_transfer".

#Question: We have the following set of intents along with their descriptions:

- * wrong_amount_of_cash_received: The intent "wrong_amount_of_cash_received" refers to expressing concerns or seeking clarification when the amount of money received during a transaction does not match the expected value.
- * fiat_currency_support: The intent "fiat_currency_support" denotes inquiries or requests for assistance regarding the handling, availability, or policies related to traditional, government-issued currencies within a service or platform.
- * supported_cards_and_currencies: The intent "supported_cards_and_currencies" pertains to inquiries regarding the types of cards accepted and the range of currencies supported by a service, focusing on clarifying compatibility and transactional possibilities.
- * exchange_via_app: The intent "exchange_via_app" pertains to inquiries or actions related to conducting currency exchanges through a mobile application.
- * exchange_charge: The intent "exchange_charge" relates to inquiries or concerns about fees incurred during the process of converting one currency to another, typically involving questions about the nature, amount, or legitimacy of the charges applied.

A user wrote the following text: "Can I exchange currency to Europe?". The intent of this text is definitely one from the five intents in the previous set. What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: The user is asking about exchanging currency to Europe which falls under the category of traditional, government-issued currencies. Therefore, the intent of the user is "fiat_currency_support".

#Question: We have the following set of intents along with their descriptions:

- * atm_support: The intent "atm_support" pertains to seeking help or information related to the functioning, issues, or usage of an ATM.
- * age_limit: The intent "age_limit" refers to inquiries or concerns regarding the minimum or maximum age requirements imposed for opening a bank account.
- * top_up_by_bank_transfer_charge: The intent "top_up_by_bank_transfer_charge" refers to inquiries or concerns about any fees associated with adding funds to an account via a bank transfer.

- * cash_withdrawal_charge: The intent "cash_withdrawal_charge" pertains to inquiries or concerns regarding fees incurred when withdrawing money from an account, typically focusing on the specifics of the charge, such as its amount, reason, or the possibility of waiving it.
- * top_up_by_card_charge: The intent "top_up_by_card_charge" typically refers to inquiries or concerns related to fees incurred when adding funds to an account or service using a credit or debit card.

A user wrote the following text: ""Do you offer services for children to have money saving experience?"". The intent of this text is definitely one from the five intents in the previous set. What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

BANKING77 - 3-shot Prompt 2: without Descriptions

#Question: We have the following set of intents: ['cancel_transfer', 'failed_transfer', 'transfer_not_received_by_recipient', 'edit_personal_details', 'request_refund'].

A user wrote the following text: "Help! I need to cancel a transaction."

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text "Help! I need to cancel a transaction", the intent of the user is to request a cancellation of a transfer, specifically a transfer that has already been initiated and needs to be stopped or reversed, which aligns with the intent "cancel_transfer" from the provided set.

#Question: We have the following set of intents: ['extra_charge_on_statement', 'card_swallowed', 'cash_withdrawal_charge', 'card_payment_fee_charged', 'declined_cash_withdrawal'].

A user wrote the following text: "What is your policy on withdrawing cash from an ATM? There is a fee on my account that wasn't there before I paid for my groceries."

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text, the intent of the user is to ask about the policy on withdrawing cash from an ATM, specifically asking about any fees associated with such a transaction, which aligns with the "cash_withdrawal_charge" intent.

#Question: We have the following set of intents: ['order_physical_card', 'age_limit', 'unable_to_verify_identity', 'verify_my_identity', 'why_verify_identity'].

A user wrote the following text: "Let me know what the steps for the identity checks are"

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer: Based on the text spoken by the user "Let me know what the steps for the identity checks are", their intent is to ask about the process or steps required for verifying their identity. Therefore, the intent of the user can be classified under "verify_my_identity" from the given set of intents.

#Question: We have the following set of intents: ['receiving_money', 'exchange_charge', 'fiat_currency_support', 'exchange_via_app', 'top_up_by_card_charge'].

A user wrote the following text: "Does it cost extra to exchange currencies?"

What was the intent of the user when they wrote this text? Let's take a deep breath and work on this step by step, answer in a single sentence, do not speculate and for your answer include the intent as written in the previous set, exactly as it is written there.

#Answer:

Correct Responses - CLINC-150

B

In this Appendix, we present a sample of correctly classified texts from CLINC-150 with their Llama responses.

method: *Show your Thoughts*

text: both could be

response: the intent of the user when they wrote the text "both could be" is "maybe" because the user is expressing uncertainty and exploring multiple possibilities.

method: *Show your Thoughts*

text: is my request going to be approved for vacation

response: the intent of the user when they wrote the text "is my request going to be approved for vacation" is "pto_request_status".

method: *Show your Thoughts*

text: can you tell me how to spell the word schadenfreude, i don't know how it's spelt

response: the intent of the user when they wrote the text "can you tell me how to spell the word schadenfreude, i don't know how it's spelt" is the "spelling" intent because the user is asking for clarification on the correct spelling of the word "schadenfreude".

method: *Show your Thoughts*

text: my credit score stinks but i don't know how to improve it

response: the intent of the user when they wrote the text "my credit score stinks but i don't know how to improve it" is "improve_credit_score". this is because the user is expressing a desire to improve their credit score, which aligns with the intent of seeking guidance or information on strategies, tips, or actions to enhance one's credit score.

method: *Show your Thoughts*

text: i was at publix trying to buy some food but my card wouldn't work

response: the intent of the user when they wrote the text "i was at publix trying to buy some food but my card wouldn't work" is most likely "card_declined" because the user is seeking clarification or assistance due to their card being declined during a transaction.

method: *Show your Thoughts*

text: do you have any monkey jokes

response: the intent of the user when they wrote the text "do you have any monkey jokes" is "tell_joke". this is because the user is seeking amusement by requesting the delivery of a humorous or entertaining anecdote, which aligns with the definition of the "tell_joke" intent.

method: *Show your Thoughts*

text: i need you to call the phone company for me

response: the intent of the user when they wrote the text "i need you to call the phone company for me" is "make_call". this is because the user is requesting that someone (the assistant) make a call to the phone company on their behalf, which aligns with the definition of the "make_call" intent.

method: *Show your Thoughts*

text: what is up with my account, why is it blocked

response: the intent of the user when they wrote the text "what is up with my account, why is it blocked" is "account_blocked". this is because the user is seeking clarification and assistance regarding the status of their account, which has been blocked or restricted from normal functionality.

method: *Show your Thoughts*

text: could you use the female voice

response: the intent of the user when they wrote the text "could you use the female voice" is "change_accent", as it pertains to a request related to modifying the voice assistant's accent or language pronunciation.

method: *Show your Thoughts*

text: i've lost my phone somewhere

response: the intent of the user when they wrote the text "i've lost my phone somewhere" is "find_phone". this is because the text explicitly states that the user has lost their phone and is seeking help in finding it, which aligns with the description of the "find_phone" intent.

method: *Show your Thoughts*

text: what is the city mpg on this car

response: the intent of the user when they wrote the text "what is the city mpg on this car" is "mpg". this is because the user is asking for information about the fuel efficiency of the car in city driving conditions, which aligns with the "mpg" intent.

method: *Show your Thoughts*

text: what's the name of the person who made you

response: the intent of the user when they wrote the text "what's the name of the person who made you" is "who_made_you", as it is listed in the previous set of intents as an inquiry seeking information about the creators or developers responsible for designing and building a particular system, application, or entity.

method: *Deep Breath*

text: can you check what the problem with my account is, it seems to be blocked

response: the intent of the user when they wrote the text "can you check what the problem with my account is, it seems to be blocked" is "account_blocked".

method: *Deep Breath*

text: open map and pin point my location

response: the intent of the user when they wrote the text "open map and pin point my location" is "current_location".

method: *Deep Breath*

text: if i use applesauce instead of oil, will that mess it up

response: the intent of the user when they wrote the text "if i use applesauce instead of oil, will that mess it up" is "ingredient_substitution".

method: *Deep Breath*

text: can you tell me what hobbies the ai has

response: the intent of the user when they wrote the text "can you tell me what hobbies the ai has" is "what_are_your_hobbies", as it matches the description of the intent "seeking information about the personal interests and activities of the ai".

method: *Deep Breath*

text: can i increase the credit limit on my mastercard

response: the intent of the user when they wrote the text "can i increase the credit limit on my mastercard" is "credit_limit_change".

method: *Deep Breath*

text: use my navy federal account and pay this month's verizon bill, please

response: the intent of the user when they wrote the text "use my navy federal account and pay this month's verizon bill, please" is "pay_bill".

method: *Deep Breath*

text: can you verify my recent visa card transaction

response: the intent of the user when they wrote the text "can you verify my recent visa card transaction" is "transactions".

method: *Deep Breath*

text: who do you report to for your work

response: the intent of the user when they wrote the text "who do you report to for your work" is "who_do_you_work_for".

method: *Deep Breath*

text: i heard some woman say she was going to yerd me, what's that mean

response: based on the text provided, the intent of the user is likely "definition" because they are asking for clarification on what a specific phrase ("yerd me") means.

method: *Deep Breath*

text: what would be the best way to deal with lost luggage

response: the intent of the user when they wrote the text "what would be the best way to deal with lost luggage" is "lost_luggage".

method: *Deep Breath*

text: how can i find out what my tire pressure is

response: the intent of the user when they wrote the text "how can i find out what my tire pressure is" is "tire_pressure".

method: *Deep Breath*

text: my answer to your question is yes

response: the intent of the user when they wrote the text "my answer to your question is yes" is "yes".

method: *Let's Think*

text: i want to talk english

response: the intent of the user when they wrote the text "i want to talk english" is "change_language" because the user is seeking guidance on how to switch to speaking english on a device or platform.

method: *Let's Think*

text: what's my yearly salary

response: the intent of the user when they wrote the text "what's my yearly salary" is "income".

method: *Let's Think*

text: are there any open reservations at outback tonight

response: the intent of the user when they wrote the text "are there any open reservations at outback tonight" is "restaurant_reservation". this is because the text indicates a desire to make a reservation at a specific dining establishment (outback) for a specific date and time (tonight).

method: *Let's Think*

text: i want to only use the whisper sound

response: the intent of the user when they wrote "i want to only use the whisper sound" is "whisper_mode".

method: *Let's Think*

text: find a decent rental care for next week at the philadelphia airport

response: the intent of the user when they wrote the text "find a decent rental care for next week at the philadelphia airport" is most likely "car_rental" because the text mentions renting a car and providing specific location and time frame information, which aligns with the description of the "car_rental" intent.

method: *Let's Think*

text: can i just use oil instead of vinegar

response: the intent of the user when they wrote the text "can i just use oil instead of vinegar" is "ingredient_substitution". this is because the user is asking if they can substitute one ingredient (oil) for another (vinegar) in a recipe.

method: *Let's Think*

text: do you know how to spell curiosity

response: the intent of the user when they wrote the text "do you know how to spell curiosity" is "spelling". this is because the user is asking for the correct spelling of the word "curiosity", which falls under the "spelling" intent category as described in the previous set of intents.

method: *Let's Think*

text: what is the information on this song

response: the intent of the user when they wrote the text "what is the information on this song" is "what_song".

method: *Let's Think*

text: how many carry ons can i take on a flight with united airlines to austin

response: the intent of the user when they wrote the text "how many carry ons can i take on a flight with united airlines to austin" is "carry_on".

method: *Let's Think*

text: could you find out if my bank charges transaction fees on my american card if use internationally

response: the intent of the user when they wrote the text "could you find out if my bank charges transaction fees on my american card if use internationally" is "international_fees".

method: *Let's Think*

text: how many stars does brasserie les halles have

response: the intent of the user when they wrote the text "how many stars does brasserie les halles have" is "restaurant_reviews". this is because the user is seeking information about the rating or reputation of a specific dining establishment, which aligns with the description of the "restaurant_reviews" intent.

method: *Let's Think*

text: what is the information on this song

response: the intent of the user when they wrote the text "what is the information on this song" is "what_song".

method: 1-shot

text: i need to know the carry-on rules for southwest

response: since the user needs to know the carry-on rules for southwest airlines, which falls under the category of seeking information about carry-on luggage, the intent is "carry_on".

method: 1-shot

text: how do i find the interest rate for my checking account

response: since the user is asking how to find the interest rate for their checking account, which falls under the category of seeking information about the current interest rates on loans or financial products, the intent is "interest_rate".

method: 1-shot

text: thanks for all your help

response: since the user is expressing gratitude and appreciation for assistance, the intent is "thank_you".

method: 1-shot

text: do you take your orders from me

response: since the user is asking if someone takes their orders from them, which is a question seeking information about the chain of command or hierarchy within an organization, the intent is "who_do_you_work_for".

method: 1-shot

text: talk fast

response: since the user requested to talk fast, which is related to changing the speaking speed, the intent is "change_speed".

method: 1-shot

text: what's the news with my delta flight

response: since the user is asking about the status of their delta flight, which falls under the category of seeking information about a specific flight, the intent is "flight_status".

method: 1-shot

text: can someone look at my check engine light that's on

response: since the user is asking for someone to look at their check engine light, which is a warning indicator on their vehicle's dashboard, the intent is "schedule_maintenance". this is because the user is likely seeking assistance with arranging maintenance for their car, specifically to diagnose and address the issue indicated by the check engine light.

method: 1-shot

text: what firm do you work for

response: since the user is asking what firm the entity they are interacting with works for, which falls under the category of seeking information about the entity's affiliation or employer, the intent is "who_do_you_work_for".

method: 1-shot

text: change name for me

response: since the user is asking to change their name, which falls under the category of modifying their username within a digital platform or application, the intent is "change_user_name".

method: 1-shot

text: can you please instruct me on the proper way to change the oil in my car

response: since the user is asking for instructions on the proper way to change the oil in their car, which falls under the category of seeking guidance on the process of performing an oil change, the intent is "oil_change_how".

method: 1-shot

text: travel alert in cuba

response: since the user is looking for information about travel alerts in cuba, which falls under the category of seeking information about potential disruptions or noteworthy events that may impact travel plans, the intent is "travel_alert".

method: 1-shot

text: what is the information on this song

response: since the user is asking about the types of things that the assistant can help them with, which falls under the category of seeking guidance or suggestions, the intent is "what_can_i_ask_you".

method: 2-shot

text: milk should be added to my shopping list

response: based on the text provided, the user's intent is to add milk to their shopping list, which aligns with the intent "shopping_list_update", as they are seeking to modify their existing shopping list by adding a new item.

method: 2-shot

text: how soon milk expires

response: based on the text provided, the user's intent is to seek information about the shelf life or expiration date of milk, which falls under the category of "food_last", as the user is inquiring about the remaining time until the milk goes bad.

method: 2-shot

text: can you please instruct me on the proper way to change the oil in my car

response: based on the text provided, the user's intent is to seek guidance or instructions on the process of performing an oil change for their car, specifically looking for the proper way to change the oil. therefore, the intent of the user is "oil_change_how".

method: 2-shot

text: is there an uber that drives to the bank on 5th that can hold 3 people

response: based on the text provided, the user's intent is to request an uber ride that can accommodate three people and drive them to a bank located on 5th street. therefore, the intent of the user is "uber".

method: 2-shot

text: should i have my shots up to date before i travel to the virgin islands

response: based on the text provided, the user's intent is to seek information about the necessary vaccinations before traveling to the virgin islands, specifically inquiring about whether they should have their shots up to date. therefore, the intent of the user is "vaccines".

method: 2-shot

text: i need to call my boss

response: based on the text provided, the user's intent is to initiate a phone call to their boss, which aligns with the intent "make_call".

method: 2-shot

text: freeze my account for me

response: based on the text provided, the user's intent is to seek guidance or initiate a request to freeze their account temporarily, therefore the intent is "freeze_account".

method: 2-shot

text: what specific health plan do i have

response: based on the text provided, the user's intent is to seek information about their specific health plan, which falls under the category of "insurance", so the intent is "insurance".

method: 2-shot

text: let me know how many pounds are in 10 kilos

response: based on the text provided, the user's intent is to seek information about the conversion of weight measurements from kilograms (kg) to pounds (lb), specifically to determine how many pounds are equivalent to 10 kilograms. therefore, the intent of the user is "measurement_conversion".

method: 2-shot

text: tell me a joke about elephants

response: based on the text provided, the user's intent is to request the delivery of a humorous or entertaining anecdote about elephants, therefore the intent is "tell_joke".

method: 2-shot

text: i wanna know how long until the next day off

response: based on the text provided, the user's intent is to seek information about the upcoming public or personal holiday, specifically asking how long until the next day off. therefore, the intent of the user is "next_holiday".

method: 2-shot

text: what's the updated status of my order

response: based on the text provided, the user's intent is to seek information about the current status of their order, which falls under the category of "order_status", as they inquire about the updated status of their purchase.

method: 3-shot

text: give me the date that i last got my car checked

response: based on the given text, the intent of the user when they wrote "give me the date that i last got my car checked" is to seek information about the last time their car was serviced or maintained, which aligns with the "last_maintenance" intent from the provided set of intents.

method: 3-shot

text: what have i spent on food this month

response: based on the text provided, the user's intent is to seek information about their past expenditures or financial transactions related to food purchases for the current month, which aligns with the "spending_history" intent from the provided set of intents.

method: 3-shot

text: how do i find the interest rate for my checking account

response: based on the text provided, the user's intent is to seek information about the interest rate associated with their checking account, which aligns with the "interest_rate" intent from the provided set of intents.

method: 3-shot

text: tell me a dog joke

response: based on the given text, the intent of the user when they wrote "tell me a dog joke" is to seek amusement by requesting the delivery of a humorous or entertaining anecdote, which aligns with the "tell_joke" intent from the provided set of intents.

method: 3-shot

text: set an alarm for me

response: based on the text provided, the user's intent is to set an alarm, which aligns with the "alarm" intent from the provided set of intents.

method: 3-shot

text: how many stars does brasserie les halles have

response: based on the given text, the intent of the user when they wrote "how many stars does brasserie les halles have" is to seek information about the rating or reputation of brasserie les halles, which aligns with the "restaurant_reviews" intent from the provided set of intents.

method: 3-shot

text: what is the expiration date for honey

response: based on the given text, the intent of the user when they wrote "what is the expiration date for honey" is to seek information about the shelf life or expiration date of honey, which aligns with the "food_last" intent from the provided set of intents.

method: 3-shot

text: do you have any idea if the flying wing does reservations

response: based on the text provided, the user's intent is to inquire about the availability of reservations at the flying wing, which aligns with the "accept_reservations" intent from the provided set of intents.

method: 3-shot

text: my car battery is dead what do i do

response: based on the text provided, the user's intent is to seek guidance or assistance in jump-starting their car, which aligns with the "jump_start" intent from the provided set of intents.

method: *3-shot*

text: roll five sided dice

response: based on the given text, the intent of the user when they wrote "roll five sided dice" is to request the outcome of rolling a five-sided die, which aligns with the "roll_dice" intent from the provided set of intents.

method: *3-shot*

text: find out where i am on the map

response: based on the text provided, the intent of the user when they wrote "find out where i am on the map" is to determine their current location, which aligns with the "current_location" intent from the provided set of intents.

method: *3-shot*

text: i am asking you to add this song to my country music playlist

response: based on the given text, the intent of the user when they wrote "i am asking you to add this song to my country music playlist" is to update their existing music playlist by adding a new song, which aligns with the "update_playlist" intent from the provided set of intents.

Correct Responses -

BANKING77

In this Appendix, we present a sample of correctly classified texts from BANKING77 with their Llama responses.

method: *Deep Breath*

text: If the transfer details have already but reviewed and confirmed that they are correct, what other reason would cause my transfer to be declined?

response: sure! based on the text the user wrote, the intent is: "declined_transfer" because the user is asking about reasons why their transfer might be declined, which aligns with the description of the "declined_transfer" intent.

method: *Deep Breath*

text: My card was taken by the ATM.

response: the intent of the user when they wrote the text "my card was taken by the atm" is "card_swallowed".

method: *Deep Breath*

text: Why was I charged for card payment?

response: the intent of the user when they wrote the text "why was i charged for card payment?" is "card_payment_fee_charged".

method: *Deep Breath*

text: I would like to know why I was charged twice for my purchase.

response: the intent of the user when they wrote this text is "transaction_charged_twice".

method: *Deep Breath*

text: How do I get the contactless feature to work for my card?

response: the intent of the user when they wrote the text "how do i get the contactless feature to work for my card?" is "contactless_not_working".

method: *Deep Breath*

text: the balance on my account didnt change when i transferred money

response: the intent of the user when they wrote the text "the balance on my account didn't change when i transferred money" is "balance_not_updated_after_bank_transfer".

method: *Deep Breath*

text: How long is the wait for a US transfer?

response: the intent of the user when they wrote the text "how long is the wait for a us transfer?" is "transfer_timing".

method: *Deep Breath*

text: What are your currency exchange fees?

response: based on the text "what are your currency exchange fees?", the user's intent is most likely "exchange_charge".

method: *Deep Breath*

text: The ATM at Metro bank on High St. Kensington didn't return my card. What should I do now that the bank is closed?

response: based on the text, the user's intent is most likely "card_swallowed" because the user is inquiring about an issue with an atm retaining their bank card after a transaction.

method: *Deep Breath*

text: Is there any documentation needed for the identity check?

response: the intent of the user when they wrote the text "is there any documentation needed for the identity check?" is "verify_my_identity".

method: *Deep Breath*

text: Where do you deliver cards by mail?

response: sure! based on the text "where do you deliver cards by mail?", the user's intent is most likely "order_physical_card". this is because the user is asking about the delivery of a physical card by mail, which aligns with the intent "order_physical_card" as it pertains to queries or actions related to requesting a new physical card.

method: *Deep Breath*

text: How do I get my paycheck through this?

response: the intent of the user when they wrote the text "how do i get my paycheck through this?" is "receiving_money".

method: *Show your Thoughts*

text: Can my 19 year old daughter open a savings account at the bank?

response: the intent of the user when they wrote the text "can my 19 year old daughter open a savings account at the bank?" is "age_limit". this is because the user is inquiring about the minimum age requirement for opening a savings account, which falls under the "age_limit" intent category.

method: *Show your Thoughts*

text: Is it possible to use Apple Pay to put money in my account?

response: the intent of the user when they wrote the text "is it possible to use apple pay to put money in my account?" is "apple_pay_or_google_pay".

method: *Show your Thoughts*

text: Why does the top-up need verification?

response: the intent of the user when they wrote the text "why does the top-up need verification?" is "verify_top_up". this is because the user is asking about the need for verification in relation to a top-up, which aligns with the description of the "verify_top_up" intent.

method: *Show your Thoughts*

text: I would like a refund for something I bought

response: the intent of the user when they wrote the text "i would like a refund for something i bought" is "request_refund".

method: *Show your Thoughts*

text: If I am overseas, how do I change my pin?

response: the intent of the user when they wrote the text "if i am overseas, how do i change my pin?" is "change_pin" because the user is asking about modifying or setting a new personal identification number for their card or account while being abroad.

method: *Show your Thoughts*

text: My pin was blocked, how do I make it so I can use it?

response: the intent of the user when they wrote the text "my pin was blocked, how do i make it so i can use it?" is "pin_blocked".

method: *Show your Thoughts*

text: can you tell me what cards and currencies you take?

response: the intent of the user when they wrote the text "can you tell me what cards and currencies you take?" is "supported_cards_and_currencies".

method: *Show your Thoughts*

text: There's a Direct Debit payment in my account that I didn't make

response: the intent of the user when they wrote the text "how old do you have to be to be able to open an account?" is "age_limit".

method: *Show your Thoughts*

text: How old do you have to be to be able to open an account?

response: the intent of the user when they wrote the text "there's a direct debit payment in my account that i didn't make" is "direct_debit_payment_not_recognized". this is because the user is indicating that there is a direct debit payment in their account that they did not make, which aligns with the description of the "direct_debit_payment_not_recognized" intent.

method: *Show your Thoughts*

text: Are there extra charges for exchanging currency?

response: the intent of the user when they wrote the text "are there extra charges for exchanging currency?" is "exchange_charge". this is because the user is inquiring about potential additional fees associated with converting one currency to another, which aligns with the description of the "exchange_charge" intent.

method: *Show your Thoughts*

text: The exchange rate you gave me for my cash withdrawal is wrong

response: based on the text provided, the intent of the user is "wrong_exchange_rate_for_cash_withdrawal". this is because the user is stating that the exchange rate given to them for a cash withdrawal is incorrect, which aligns with the description of the "wrong_exchange_rate_for_cash_withdrawal" intent.

method: *Show your Thoughts*

text: Where do I go if I want to link my new card?

response: the intent of the user when they wrote the text "where do i go if i want to link my new card?" is "card_linking".

method: *Let's Think*

text: Is there a discount for frequently exchanging currencies?

response: the intent of the user when they wrote the text "is there a discount for frequently exchanging currencies?" is "exchange_charge".

method: *Let's Think*

text: Is it acceptable to use my card anywhere?

response: the intent of the user when they wrote the text "is it acceptable to use my card anywhere?" is "card_acceptance". this is because the user is asking about the acceptance of their card at various establishments, which aligns with the description of the "card_acceptance" intent.

method: *Let's Think*

text: I tried to get \$100 but I just got \$20

response: the intent of the user when they wrote the text "i tried to get \$100 but i just got \$20" is "wrong_amount_of_cash_received".

method: *Let's Think*

text: I purchased some makeup through a site in China, and I was under the impression that when I make transfers there is no fee. Why am I seeing this fee now? I am not happy about this at all.

response: the intent of the user when they wrote this text is "transfer_fee_charged". the user is inquiring about unexpected fees being charged for a transfer, specifically mentioning that they were under the impression there would be no fee.

method: *Let's Think*

text: How can I transfer funds from my bank to my top up account?

response: the intent of the user when they wrote the text "how can i transfer funds from my bank to my top up account?" is "transfer_into_account".

method: *Let's Think*

text: Why did it decline my payment?

response: the intent of the user when they wrote the text "why did it decline my payment?" is "declined_card_payment". this is because the text explicitly mentions a declined payment, which aligns with the description of the "declined_card_payment" intent.

method: *Let's Think*

text: places i can use the card

response: the intent of the user when they wrote the text "places i can use the card" is "card_acceptance". this is because the text is asking for information about where the card can be used, which falls under the category of card acceptance.

method: *Let's Think*

text: lost my phone, dont want others to use it.

response: the intent of the user when they wrote the text "lost my phone, don't want others to use it" is "lost_or_stolen_phone".

method: *Let's Think*

text: Besides USD what other currencies can I have?

response: the intent of the user when they wrote the text "besides usd what other currencies can i have?" is "fiat_currency_support".

method: *Let's Think*

text: Where will I find my card?

response: the intent of the user when they wrote the text "where will i find my card?" is "order_physical_card". this is because the text is asking about finding a physical card, which aligns with the intent of requesting a new physical card.

method: *Let's Think*

text: I can't find my phone but it has all my credit card information.

response: the intent of the user when they wrote the text "i can't find my phone but it has all my credit card information" is "lost_or_stolen_phone".

method: *Let's Think*

text: How can I lookup where funds came from?

response: the intent of the user when they wrote the text "how can i lookup where funds came from?" is "verify_source_of_funds".

method: *1-shot*

text: Why doesn't my balance reflect my transfer

response: based on the text "why doesn't my balance reflect my transfer", the user is inquiring about the reason why their account balance does not reflect the recent transfer they made, and therefore the intent of the user is "balance_not_updated_after_bank_transfer".

method: *1-shot*

text: Where can i find your locations?

response: based on the text "where can i find your locations?", the user is seeking information about the physical locations where the services or products offered by the entity are available, and therefore the intent of the user is "country_support".

method: *1-shot*

text: How can i transfer money from an outside bank?

response: based on the text "how can i transfer money from an outside bank?", the user is seeking assistance or guidance on how to transfer funds from an external bank account into their current account or another destination, therefore the intent of the user is "transfer_into_account".

method: *1-shot*

text: Why am i required to verify my identity?

response: based on the text "why am i required to verify my identity?", the user is seeking clarification and understanding about the reasons behind the need for identity verification procedures, therefore the intent of the user is "why_verify_identity".

method: *1-shot*

text: Does delivery to the us take long?

response: based on the text "does delivery to the us take long?", the user is inquiring about the estimated delivery time for a card to the united states, specifically asking about the duration of the delivery process. therefore, the intent of the user is "card_delivery_estimate".

method: *1-shot*

text: Where can i go to modify my detail?

response: based on the text "where can i go to modify my detail?", the user is seeking assistance in updating their personal information, therefore the intent of the user is "edit_personal_details".

method: *1-shot*

text: Pending transfer?

response: based on the text "pending transfer?", the user is inquiring about the status of a financial transaction that has been initiated but not yet completed, therefore the intent of the user is "pending_transfer".

method: *1-shot*

text: Why doesn't the app believe i am who i say i am?

response: based on the text "why doesn't the app believe i am who i say i am?", the user is seeking clarification and understanding about the reasons behind the app's inability to recognize them as who they claim to be, therefore the intent of the user is "unable_to_verify_identity".

method: *1-shot*

text: If i need more cards, are there any fees?

response: based on the text "if i need more cards, are there any fees?", the user is seeking information about potential fees associated with acquiring additional cards, and therefore the intent of the user is "getting_spare_card".

method: *1-shot*

text: I can't find my phone but it has all my credit card information

response: based on the text "i can't find my phone but it has all my credit card information", the user is seeking assistance and support for a lost or stolen phone containing sensitive financial information, therefore the intent of the user is "lost_or_stolen_phone".

method: *1-shot*

text: Can i add funds to the card directly from my bank account?

response: based on the text "can i add funds to the card directly from my bank account?", the user is inquiring about the possibility of transferring funds directly from their bank account into their card, and therefore the intent of the user is "transfer_into_account".

method: *1-shot*

text: How come the atm gave me less cash than what i asked for?

response: based on the text "how come the atm gave me less cash than what i asked for?", the user is seeking clarification and understanding regarding a discrepancy between the amount of money requested and the amount dispensed by an automated teller machine (atm), which aligns with the intent "wrong_amount_of_cash_received".

method: 2-shot

text: Where are the locations of atms that accept this card?

response: based on the text "where are the locations of atms that accept this card?", the user is seeking information about the location of atms that accept their specific bank card. therefore, the intent of the user is "atm_support".

method: 2-shot

text: What is this \$1 charge on my statement?

response: based on the text "what is this \$1 charge on my statement?", the user is seeking clarification and understanding about an additional charge appearing on their financial statement. therefore, the intent of the user is "extra_charge_on_statement".

method: 2-shot

text: Where's the verification code for the top-up card?

response: based on the text "where's the verification code for the top-up card?", the user is seeking confirmation of the addition of funds to their account or service, specifically asking for the verification code related to the top-up card. therefore, the intent of the user is "verify_top_up".

method: 2-shot

text: I have withdrawn cash from atm but i received the wrong amount. i want cash back as in app its showing actual amount which i got. please help me in this.

response: based on the text "i have withdrawn cash from atm but i received the wrong amount. i want cash back as in app its showing actual amount which i got. please help me in this.", the user is seeking assistance for a wrong amount of cash received during an atm withdrawal, specifically requesting cash back. therefore, the intent of the user is "wrong_amount_of_cash_received".

method: 2-shot

text: Why didn't the atm give me my money?

response: based on the text "why didn't the atm give me my money?", the user is seeking clarification and resolution for an unsuccessful attempt to withdraw cash from an account. therefore, the intent of the user is "declined_cash_withdrawal".

method: 2-shot

text: When will the transfer be completed?

response: based on the text "when will the transfer be completed?", the user is seeking information about the expected completion time of a financial transaction. therefore, the intent of the user is "pending_transfer".

method: 2-shot

text: Can my friends send me money?

response: based on the text "can my friends send me money?", the user is seeking information about receiving money from others. therefore, the intent of the user is "receiving_money".

method: 2-shot

text: Where do i go to get the code to verify the top up card?

response: based on the text "where do i go to get the code to verify the top up card?", the user is seeking information on how to obtain a verification code to confirm the addition of funds to their account. therefore, the intent of the user is "verify_top_up".

method: 2-shot

text: Please, activate my card,

response: based on the text "please, activate my card," the user is requesting to activate their card, which aligns with the intent "activate_my_card."

method: 2-shot

text: My card is about to expire,

response: based on the text "my card is about to expire," the user is seeking information or assistance regarding the upcoming expiration of their payment card. therefore, the intent of the user is "card_about_to_expire."

method: 2-shot

text: I need help with a lost card,

response: Based on the text "i need help with a lost card," the user is seeking assistance related to a lost or stolen card, specifically regarding the security and potential fraudulent activities associated with it. therefore, the intent of the user is "lost_or_stolen_card."

method: 2-shot

text: Hello. i'm on holiday and didn't bring any cash with me. i need to withdrawal my home currency from one of your machines. do you have any that will do this and is there a charge?

response: based on the text "hello. i'm on holiday and didn't bring any cash with me. i need to withdrawal my home currency from one of your machines. do you have any that will do this and is there a charge?", the user is seeking information about withdrawing their home currency from an atm and if there are any charges associated with it. therefore, the intent of the user is "wrong_exchange_rate_for_cash_withdrawal".

method: 3-shot

text: Do you offer services for children to have money saving experience?

response: the user is asking if the service offers experiences for children to learn about money management, which falls under the category of age limits and requirements. therefore, the intent of the user is "age_limit".

method: 3-shot

text: Can I cancel a transfer if is already processed?

response: the user is asking if they can cancel a transfer that has already been processed, which falls under the category of attempting to halt or reverse a previously initiated monetary transaction. therefore, the intent of the user is "cancel_transfer".

method: 3-shot

text: Do you know why my card payment is declined?

response: the user is asking about the reason for the decline of their card payment, which falls under the category of "declined_card_payment". therefore, the intent of the user is "declined_card_payment".

method: 3-shot

text: How can I get paid in a different currency?

response: the user is asking how to receive payment in a different currency, which falls under the category of "receiving money." therefore, the intent of the user is "receiving_money".

method: 3-shot

text: How long will it take to arrive?

response: the user is asking about the estimated time of arrival of a card, which falls under the category of "card_delivery_estimate". therefore, the intent of the user is "card_delivery_estimate".

method: 3-shot

text: I'd like to have another card

response: based on the text "i'd like to have another card," the user is seeking to acquire an additional card, which falls under the category of "getting_spare_card." therefore, the intent of the user is "getting_spare_card."

method: 3-shot

text: I think something may have happened that caused a charge to show up twice.

response: based on the text "i think something may have happened that caused a charge to show up twice," the user is seeking clarification and resolution regarding a double charge on their account. therefore, the intent of the user is "transaction_charged_twice."

method: 3-shot

text: How come the ATM gave me less cash than what I asked for?

response: the user is asking about receiving less cash than requested from an atm, which falls under the category of "wrong_amount_of_cash_received". therefore, the intent of the user is "wrong_amount_of_cash_received".

method: 3-shot

text: Can my 19 year old daughter open a savings account at the bank?

response: based on the text "can my 19 year old daughter open a savings account at the bank?", the user is seeking information about the eligibility criteria for opening a savings account, specifically regarding the minimum age requirement. therefore, the intent of the user is "age_limit".

method: 3-shot

text: Is it possible to make several disposable cards in a day?

response: the user is asking if it is possible to create multiple disposable cards in a day, which falls under the category of "disposable_card_limits." therefore, the intent of the user is "disposable_card_limits."

method: 3-shot

text: why hasn't my top up gone through yet

response: based on the text "why hasn't my top up gone through yet," the user is inquiring about the status of a funds addition transaction that has not yet been processed or credited to their account. therefore, the intent of the user is "pending_top_up."

method: 3-shot

text: what is the identity check?

response: the user is asking about the identity check, which falls under the category of verifying personal identification. therefore, the intent of the user is "why_verify_identity".

Error Analysis Contents -

CLINC-150

In this Appendix, we present the samples we collected for the error analysis of CLINC-150.

text: is it possible to make reservations at the new york restaurant

intent: accept_reservations

response: based on the text provided, the user's intent is to make a reservation at the new york restaurant, which aligns with the "restaurant_reservation" intent from the given set of intents.

predicted intent : restaurant_reservation

Error Type: Correct Chain of Thought, Wrong Classification

text: do i have money in my wells fargo account for nike's

intent: balance

response: based on the text provided, the user's intent is to inquire about the current balance of their wells fargo account to determine if they have enough funds to purchase nike's products, which aligns with the "bill_balance" intent from the provided set of intents.

predicted intent : bill_balance

Error Type: Correct Chain of Thought, Wrong Classification

text: you need to shut it

intent: cancel

response: based on the text provided, the user's intent is to express a negative response or refusal, which aligns with the "no" intent from the previous set of intents.

predicted intent : no

Error Type: Wrong Chain of Thought

text: please go ahead and change my name

intent: change_user_name

response: based on the text provided, the user's intent is to request a change to their name, which aligns with the "change_ai_name" intent from the provided set of intents.

predicted intent : change_ai_name

Error Type: Correct Chain of Thought, Wrong Classification

text: where is the nearest circuit city

intent: directions

response: based on the given text, the intent of the user when they wrote "where is the nearest circuit city" is to seek information about the current location of a circuit city store, which aligns with the "current_location" intent from the provided set of intents.

predicted intent : current_location

Error Type: Wrong Chain of Thought

text: please put my account on hold

intent: freeze_account

response: based on the text provided, the user's intent is to request that their account be placed on hold, which aligns with the "account_blocked" intent from the provided set of intents.

predicted intent : account_blocked

Error Type: Correct Chain of Thought, Wrong Classification

text: it was a pleasure talking to you

intent: goodbye

response: based on the text provided, the user's intent is to express gratitude and goodwill, which aligns with the "thank_you" intent from the provided set of intents.

predicted intent : thank_you

Error Type: Correct Chain of Thought, Wrong Classification

text: it was nice to chat

intent: goodbye

response: based on the text provided, the user's intent is to express goodwill and politeness towards the recipient, indicating a friendly and respectful tone, which aligns with the "thank_you" intent from the provided set of intents.

predicted intent : thank_you

Error Type: Correct Chain of Thought, Wrong Classification

text: what ingredients are necessary to make spaghetti carbonara

intent: ingredients_list

response: based on the given text, the user's intent is to seek information about the ingredients required to prepare spaghetti carbonara, which aligns with the "recipe" intent from the provided set of intents.

predicted intent : recipe

Error Type: Correct Chain of Thought, Wrong Classification

text: what do i need to make pasta

intent: ingredients_list

response: based on the given text, the intent of the user when they wrote "what do i need to make pasta" is to seek information about the ingredients required to prepare pasta, which aligns with the "recipe" intent from the provided set of intents.

predicted intent : recipe

Error Type: Correct Chain of Thought, Wrong Classification

text: i really would like to get a new insurance plan

intent: insurance_change

response: based on the text provided, the intent of the user when they wrote "i really would like to get a new insurance plan" is to seek information or guidance related to obtaining a new insurance policy, which aligns with the "insurance" intent from the provided set of intents.

predicted intent : insurance

Error Type: Correct Chain of Thought, Wrong Classification

text: what do i need to do to get new insurance

intent: insurance_change

response: based on the text provided, the user's intent is to seek information or guidance related to obtaining new insurance coverage, which aligns with the "insurance" intent from the provided set of intents.

predicted intent : insurance

Error Type: Correct Chain of Thought, Wrong Classification

text: are there meetings scheduled for today

intent: meeting_schedule

response: based on the text provided, the user's intent is to seek information about the current level of activity or occupancy, specifically regarding meetings scheduled for today, which aligns with the "how_busy" intent from the provided set of intents.

predicted intent : how_busy

Error Type: Wrong Chain of Thought

text: where can i find information on oil changes for my car

intent: oil_change_when

response: based on the text provided, the user's intent is to seek information about oil changes for their car, which aligns with the "oil_change_how" intent from the provided set of intents.

predicted intent : oil_change_how

Error Type: Correct Chain of Thought, Wrong Classification

text: is there a status update on my vacation approval

intent: pto_request_status

response: based on the text provided, the user's intent is to seek information about the current status of their vacation approval, specifically requesting an update on its status. this aligns with the "application_status" intent from the given set of intents.

predicted intent : application_status

Error Type: Correct Chain of Thought, Wrong but Plausible Classification

text: i need a reminder to give the dog his medicine at ten tonight

intent: reminder_update

response: based on the text provided, the user's intent is to set a reminder for giving their dog medication at 10 pm, which aligns with the "reminder" intent from the provided set of intents.

predicted intent : reminder

Error Type: Correct Chain of Thought, Wrong but Plausible Classification

text: make a reminder to pay the mortgage

intent: reminder_update

response: based on the given text, the intent of the user when they wrote "make a reminder to pay the mortgage" is to request assistance in setting up a reminder for a specific task or event, specifically to remember to make a payment, which aligns with the "reminder" intent from the provided set of intents.

predicted intent : reminder

Error Type: Correct Chain of Thought, Wrong but Plausible Classification

text: my report card was stolen

intent: report_lost_card

response: based on the given text, the intent of the user when they wrote "my report card was stolen" is to report a lost or stolen item, specifically a report card, which aligns with the "lost_luggage" intent from the provided set of intents.

predicted intent : lost_luggage

Error Type: Wrong Chain of Thought

text: my report card was stolen

intent: report_lost_card

response: based on the given text, the intent of the user when they wrote "my report card was stolen" is to report a lost or stolen item, specifically a report card, which aligns with the "lost_luggage" intent from the provided set of intents.

predicted intent : lost_luggage

Error Type: Wrong Chain of Thought

text: are there any open reservations at outback tonight

intent: restaurant_reservation

response: based on the text provided, the user's intent is to seek verification or assurance regarding open reservations at outback steakhouse for tonight, which aligns with the "confirm_reservation" intent from the given set of intents.

predicted intent : confirm_reservation

Error Type: Correct Chain of Thought, Wrong Classification

text: are there any open reservations at outback tonight

intent: restaurant_reservation

response: based on the text provided, the user's intent is to seek verification or assurance regarding open reservations at outback steakhouse for tonight, which aligns with the "confirm_reservation" intent from the given set of intents.

predicted intent : confirm_reservation

Error Type: Correct Chain of Thought, Wrong Classification

text: what do my visa points look like

intent: rewards_balance

response: based on the text provided, the user's intent is to seek information about their accumulated rewards points or benefits associated with their visa account, which aligns with the "redeem_rewards" intent from the given set of intents.

predicted intent : redeem_rewards

Error Type: Correct Chain of Thought, Wrong Classification

text: find a place near me to have my tires inspected

intent: schedule_maintenance

response: based on the text provided, the intent of the user when they wrote "find a place near me to have my tires inspected" is to seek direction or information on navigating to a specific location for their vehicle's tire inspection, which aligns with the "directions" intent from the provided set of intents.

predicted intent : directions

Error Type: Correct Chain of Thought, Wrong but Plausible Classification

text: have i already got eggs down on the shopping list

intent: shopping_list

response: based on the text provided, the user's intent is to check if eggs are already included in their shopping list, which aligns with the "shopping_list_update" intent from the provided set of intents.

predicted intent : shopping_list_update

Error Type: Correct Chain of Thought, Wrong Classification

text: have i already got eggs down on the shopping list

intent: shopping_list

response: based on the text provided, the user's intent is to check if eggs are already included in their shopping list, which aligns with the "shopping_list_update" intent from the provided set of intents.

predicted intent : shopping_list_update

Error Type: Correct Chain of Thought, Wrong Classification

text: i don't want you connected to my phone anymore

intent: sync_device

response: based on the text provided, the user's intent is to seek assistance in disconnecting or dissociating their phone from the entity they are interacting with, which aligns with the "reset_settings" intent from the provided set of intents.

predicted intent : reset_settings

Error Type: Wrong Chain of Thought

text: what time is it in phoenix

intent: timezone

response: based on the text provided, the user's intent is to seek information about the current time in phoenix, arizona, which aligns with the "time" intent from the provided set of intents.

predicted intent : time

Error Type: Correct Chain of Thought, Wrong but Plausible Classification

text: can you help me pick some fun things to do in london during my trip

intent: travel_suggestion

response: based on the given text, the intent of the user when they wrote "can you help me pick some fun things to do in london during my trip" is to seek recommendations or advice for entertaining activities and experiences in london, which aligns with the "restaurant_suggestion" intent from the provided set of intents.

predicted intent : restaurant_suggestion

Error Type: Correct Chain of Thought, Wrong Classification

text: what do you call me

intent: user_name

response: based on the given text, the intent of the user when they wrote "what do you call me" is to ask for the name by which the ai assistant refers to them, which aligns with the "what_is_your_name" intent from the provided set of intents.

predicted intent : what_is_your_name

Error Type: Correct Chain of Thought, Wrong Classification

text: say my name

intent: user_name

response: based on the text provided, the user's intent is to request the ai assistant to say its name, which aligns with the "what_is_your_name" intent from the provided set of intents.

predicted intent : what_is_your_name

Error Type: Wrong Chain of Thought

Error Analysis Contents -

BANKING77

In this Appendix we present the samples we collected for the error analysis of BANKING77.

text: Can I reactivate a card I thought I lost?

intent: card_linking

response: the user is asking if they can reactivate a card they thought they lost, which falls under the category of lost or stolen cards. therefore, the intent of the user is "lost_or_stolen_card".

predicted intent : lost_or_stolen_card

Error Type: Wrong Chain of Thought

text: Where can my card be delivered?

intent: order_physical_card

response: the user is asking about the delivery location of their card, which falls under the category of card acceptance. therefore, the intent of the user is "card_acceptance".

predicted intent : card_acceptance

Error Type: Wrong Chain of Thought

text: Where can I receive my card?

intent: order_physical_card

response: the user is asking about receiving their card, which falls under the category of getting a spare card. therefore, the intent of the user is "getting_spare_card".

predicted intent : getting_spare_card

Error Type: Wrong Chain of Thought

text: The rate applied to my foreign purchase was incorrect

intent: card_payment_wrong_exchange_rate

response: based on the text "the rate applied to my foreign purchase was incorrect," the user is seeking assistance with an issue related to an incorrect exchange rate being applied to a foreign purchase. therefore, the intent of the user is "exchange_rate."

predicted intent : exchange_rate

Error Type: Correct Chain of Thought, Wrong Classification

text: Which cash machines will allow me to change my PIN?

intent: change_pin

response: based on the text "which cash machines will allow me to change my pin?", the user is seeking information about atms that support changing their pin. therefore, the intent of the user is "atm_support".

predicted intent : atm_support

Error Type: Correct Chain of Thought, Wrong Classification

text: Where will I find my card?

intent: order_physical_card

response: based on the text "where will i find my card?", the user is seeking information about locating their physical card. therefore, the intent of the user is "getting_spare_card".

predicted intent : getting_spare_card

Error Type: Wrong Chain of Thought

text: I payed with a card and was charged an extra fee

intent: card_payment_fee_charged

response: based on the text "i paid with a card and was charged an extra fee," the user is seeking clarification and understanding about an unexpected fee charged to their card payment transaction. therefore, the intent of the user is "transfer_fee_charged."

predicted intent : transfer_fee_charged

Error Type: Correct Chain of Thought, Wrong Classification

text: I'm still waiting on my card to be delivered.

intent: card_arrival

response: based on the text "i'm still waiting on my card to be delivered," the user is seeking information or assistance regarding the status of their card delivery. therefore, the intent of the user is "card_delivery_estimate."

predicted intent : card_delivery_estimate

Error Type: Correct Chain of Thought, Wrong but Plausible Classification

text: How do I get my PIN unlocked?

intent: pin_blocked

response: the user is seeking assistance to unlock their pin, which falls under the category of regaining access to a secured account or device. therefore, the intent of the user is "passcode_forgotten".

predicted intent : passcode_forgotten

Error Type: Wrong Chain of Thought

text: I live in the US but want to get a card

intent: country_support

response: based on the text "i live in the us but want to get a card," the user is seeking information about obtaining a card, which falls under the category of "supported_cards_and_currencies." therefore, the intent of the user is "supported_cards_and_currencies."

predicted intent : supported_cards_and_currencies

Error Type: Wrong label

text: Is my PIN located on my account somewhere?

intent: get_physical_card

response: the user is asking about the location of their pin number, which is related to the card linking and activation of their card. therefore, the intent of the user is "card_linking".

predicted intent : card_linking

Error Type: Wrong label

text: My card is just not working at this time.

intent: virtual_card_not_working

response: based on the text "my card is just not working at this time," the user is experiencing issues with their card not functioning as expected, which aligns with the intent "card_not_working."

predicted intent : card_not_working

Error Type: Controversial Initial Intent

text: I couldn't do a transfer to an account

intent: beneficiary_not_allowed

response: based on the text "i couldn't do a transfer to an account," the user is seeking assistance with a failed or declined transfer. therefore, the intent of the user is "declined_transfer."

predicted intent : declined_transfer

Error Type: Wrong label

text: Can I receive a new card while I am in China?

intent: card_about_to_expire

response: the user is asking if they can receive a new card while they are in china, which falls under the category of "card_delivery_estimate". therefore, the intent of the user is "card_delivery_estimate".

predicted intent : card_delivery_estimate

Error Type: Wrong label

text: The exchange rate you gave me for my cash withdrawal is wrong

intent: wrong_exchange_rate_for_cash_withdrawal

response: based on the text "the exchange rate you gave me for my cash withdrawal is wrong," the user is expressing frustration with the incorrect exchange rate being applied to their cash withdrawal. therefore, the intent of the user is "card_payment_wrong_exchange_rate."

predicted intent : card_payment_wrong_exchange_rate

Error Type: Correct Chain of Thought, Wrong Classification

text: How do I get my PIN unlocked?

intent: pin_blocked

response: the user is seeking assistance to unlock their pin, which falls under the category of "passcode_forgotten". therefore, the intent of the user is "passcode_forgotten".

predicted intent : passcode_forgotten

Error Type: Wrong Chain of Thought

text: Can I check somewhere if there will be a fee on my payment or not? Seems you are charging in some cases, what's the pattern there?

intent: card_payment_fee_charged

response: the user is asking about the possibility of fees being charged on their payment and seeking clarification on the pattern of such charges. therefore, the intent of the user is "extra_charge_on_statement".

predicted intent : extra_charge_on_statement

Error Type: Correct Chain of Thought, Wrong Classification

text: is my card PIN saved in the app

intent: get_physical_card

response: based on the text "is my card pin saved in the app," the user is seeking information about whether their card pin is saved in the app. therefore, the intent of the user is "pin_blocked."

predicted intent : pin_blocked

Error Type: Wrong label

text: Why can't I use my virtual card for subscription services?

intent: virtual_card_not_working

response: the user is asking why they cannot use their virtual card for subscription services, which falls under the category of "getting_virtual_card" intent.

predicted intent : getting_virtual_card

Error Type: Correct Chain of Thought, Wrong Classification

text: There's a debit on my account that I didn't do.

intent: direct_debit_payment_not_recognised

response: the user is indicating that there is a debit on their account that they did not make, which falls under the category of "cash_withdrawal_not_recognised". therefore, the intent of the user is "cash_withdrawal_not_recognised".

predicted intent : cash_withdrawal_not_recognised

Error Type: Wrong Chain of Thought

text: Why does my account not accept cash deposits?

intent: balance_not_updated_after_cheque_or_cash_deposit

response: the user is asking why their account does not accept cash deposits, which falls under the category of "cash_withdrawal_not_recognized". therefore, the intent of the user is "cash_withdrawal_not_recognized".

predicted intent : cash_withdrawal_not_recognised

Error Type: Wrong Chain of Thought

text: Can I add funds to the card directly from my bank account?

intent: transfer_into_account

response: the user is asking if they can add funds directly from their bank account to their card, which falls under the category of "topping up by card." therefore, the intent of the user is "topping_up_by_card".

predicted intent : topping_up_by_card

Error Type: Wrong Chain of Thought

text: I saw a payment i did not do

intent: card_payment_not_recognised

response: based on the text "i saw a payment i did not do," the user is likely seeking assistance with a situation where an unauthorized payment was made using their card. therefore, the intent of the user is "compromised_card."

predicted intent : compromised_card

Error Type: Correct Chain of Thought, Wrong Classification

text: The exchange rate you gave me for my cash withdrawal is wrong

intent: wrong_exchange_rate_for_cash_withdrawal

response: based on the text "the exchange rate you gave me for my cash withdrawal is wrong," the user is expressing frustration with the incorrect exchange rate being applied to their cash withdrawal. therefore, the intent of the user is "card_payment_wrong_exchange_rate."

predicted intent : card_payment_wrong_exchange_rate

Error Type: Correct Chain of Thought, Wrong Classification

text: Think someone has took money out with my card. What shall I do?

intent: cash_withdrawal_not_recognised

response: the user who wrote the text "think someone has took money out with my card. what shall i do?" is likely experiencing issues related to compromised card usage. therefore, the intent of the user is "compromised_card."

predicted intent : compromised_card

Error Type: Correct Chain of Thought, Wrong Classification

text: I am not being recognized by the app.

intent: unable_to_verify_identity

response: based on the text "i am not being recognized by the app," the user is experiencing issues with the app not recognizing them, which likely relates to concerns about the security or unauthorized use of their card. therefore, the intent of the user is "compromised_card."

predicted intent : compromised_card

Error Type: Wrong Chain of Thought

text: Who else can top up my account

intent: topping_up_by_card

response: based on the text "who else can top up my account," the user is seeking information about who else can add funds to their account. therefore, the intent of the user is "transfer_into_account."

predicted intent : transfer_into_account

Error Type: Correct Chain of Thought, Wrong Classification

text: How do I show this stupid system that this is really my identity?

intent: unable_to_verify_identity

response: the user is expressing frustration and seeking assistance to prove their identity to the system. therefore, the intent of the user is "why_verify_identity".

predicted intent : why_verify_identity

Error Type: Correct Chain of Thought, Wrong Classification

text: I've attempted to do a very standard transfer and have tried 5 times at this point. Can you tell me what the issue is? Is the system down?

intent: failed_transfer

response: the user is asking about the issue with a standard transfer and if the system is down, which falls under the category of "pending_transfer". therefore, the intent of the user is "pending_transfer".

predicted intent : pending_transfer

Error Type: Wrong Chain of Thought

text: There's a cash withdraw on my statement that I didn't make.

intent: cash_withdrawal_not_recognised

response: the user is reporting an unauthorized cash withdrawal on their statement, which falls under the category of "compromised_card". therefore, the intent of the user is "compromised_card".

predicted intent : compromised_card

Error Type: Wrong Chain of Thought

List of Figures

1.1	A few-shot prompt example with Chain of Thought prompting compared with normal Prompting. Source: Wei et al. (2022)	2
3.1	Violin plot of class counts for BANKING77	10
3.2	The fine-tuning process MetaAI used on Llama. Source: Touvron et al. (2023) . .	12
4.1	An overview of the pipeline used in this thesis	13
4.2	The two-stage prompting method of 0-shot Chain of Thought prompting. Source: Kojima et al. (2022)	17
6.1	A N-shot prompt and the parts that comprise it.	32
6.2	A heatmap of all the results, both individual runs and average accuracy, using the description strategy for CLINC-150. We can easily discern between the best and worst runs and methods with ease and it becomes apparent that the best method in terms of accuracy is the 3-shot method.	35
6.3	A heatmap of all the results, both individual runs and average accuracy, using the description strategy for BANKING77. We can easily discern between the best and worst runs and methods with ease and it becomes apparent that the best method in terms of accuracy is the 3-shot method, even by a very slight margin.	41
7.1	The difference in average accuracy for CLINC-150 between using the descriptions and not using them. It is evident that the technique that profits the most by the intent descriptions is <i>Deep Breath</i> , while the few-shot methods seem to not suffer as much.	50
7.2	The difference in average accuracy for BANKING77 between using the descriptions and not using them. It is evident that the technique that profits the most by the intent descriptions is once more <i>Deep Breath</i> , while the few-shot methods seem to not hurt as much.	53

List of Tables

3.1	Statistics in both Character Level and Word Level for CLINC-150. The mean lengths are shown along with their standard deviation and the minimum, maximum, and median lengths.	8
3.2	Examples of text utterances from CLINC-150 and their assigned intent class. . .	9
3.3	Examples of text utterances from BANKING77 and their assigned intent class. .	10
3.4	Statistics in both Character Level and Word Level for BANKING77. The mean lengths are shown along with their standard deviation and the minimum, maximum, and median lengths.	10
5.1	Examples of mistakes made by Llama on inference and how we classified them.	23
5.2	Accuracy of CLINC-150 across 3 runs and their average.	23
5.3	Accuracy of BANKING77 across 3 runs and their average.	27
6.1	Accuracy of CLINC-150 across 3 runs and their average.	33
6.2	Accuracy of BANKING77 across 3 runs and their average.	38
7.1	Accuracy of CLINC-150 across 3 runs and their average. The average accuracy of the method without a Chain of Thought is highlighted.	47
7.2	Accuracy of BANKING77 across 3 runs and their average. The average accuracy of the method without a Chain of Thought is highlighted.	47
7.3	Accuracy of CLINC-150 without the use of intent descriptions. The average accuracy of the best method without intent descriptions is highlighted.	49
7.4	Accuracy of BANKING77 without the use of intent descriptions. The average accuracy of the best method without intent descriptions is highlighted.	51
7.5	Accuracy of CLINC-150 without the use of intent descriptions, <i>No CoT</i> included (highlighted).	54
7.6	Accuracy of BANKING77 without the use of intent descriptions, <i>No CoT</i> included (highlighted).	55