ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

School of Information Sciences and Technology
Department of Informatics
Athens, Greece

Bachelor Thesis
in
Computer Science

Improvements to the explanations of Ithaca's chronological attributions of ancient Greek inscriptions

Maria Schoinaki

Supervisors: Ion Androutsopoulos
Department of Informatics
Athens University of Economics and Business

John Pavlopoulos
Department of Informatics
Athens University of Economics and Business

Yannis Assael
Google DeepMind

September 2025

Maria Schoinaki

Improvements to the explanations of Ithaca's chronological attributions of ancient Greek inscriptions

September 2025

Supervisors: Ion Androutsopoulos, John Pavlopoulos, Yannis Assael

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

Transformer-based models, such as Ithaca, have proven effective in dating ancient Greek inscriptions. However, their chronological predictions remain challenging for scholars to interpret and trust. This thesis enhances Ithaca's chronological-attribution component by integrating and systematically comparing post-hoc explainability methods applied to both character-level and word-level embeddings. We introduce a unified saliency-processing pipeline that normalizes and fuses explanations into clear, token-wise heatmaps. While preserving Ithaca's original dating accuracy, our refined explanations consistently spotlight historically relevant names and terms. The accompanying code and methodological framework enhance the interpretability of Ithaca's chronological attributions, helping historians validate and understand the model's decisions.

Περίληψη

Τα μοντέλα τύπου Transformers, όπως το Ithaca, έχουν δείξει αξιοσημείωτη ικανότητα χρονολογικού προσδιορισμού αρχαίων ελληνικών επιγραφών. Παρ' όλα αυτά, οι προβλέψεις τους παραμένουν αδιαφανείς και δύσκολες στην ερμηνεία από ιστορικούς και επιγραφολόγους. Στόχος της παρούσας πτυχιακής εργασίας είναι η βελτίωση των μηχανισμών εξηγήσεων του Ithaca στην εργασία χρονολόγησης, μέσω της ενσωμάτωσης και της συστηματικής σύγκρισης μεθόδων εκ των υστέρων επεξηγησιμότητας (post-hoc explainability), οι οποίες εφαρμόζονται τόσο σε επίπεδο αναπαράστασης χαρακτήρων όσο και σε επίπεδο λέξεων.

Αναπτύσσουμε μια ενοποιημένη ροή επεξεργασίας σημαντικότητας (saliency), η οποία κανονικοποιεί τα μεμονωμένα αποτελέσματα, τα συγχωνεύει και παράγει απεικονίσεις θερμικών χαρτών (saliency maps) σε επίπεδο λέξης ή χαρακτήρα. Η προσέγγιση αυτή διατηρεί αναλλοίωτη την ακρίβεια χρονολόγησης του Ithaca, ενώ παράλληλα αναδεικνύει με συνέπεια ιστορικά κρίσιμους όρους και ονόματα που υποστηρίζουν τις χρονολογικές προβλέψεις.

Ο συνοδευτικός κώδικας και το μεθοδολογικό πλαίσιο ενισχύουν την επεξηγησιμότητα των χρονολογικών αποδόσεων του Ithaca, καθιστώντας δυνατή τη βαθύτερη κατανόηση και την επικύρωση των αποφάσεων του μοντέλου από ιστορικούς και επιγραφολόγους.

Acknowledgements

This thesis marks the culmination of my undergraduate journey at the Department of Informatics, Athens University of Economics and Business (AUEB), and I am deeply grateful to all those who have guided and supported me along the way.

First and foremost, I would like to express my deepest gratitude to my supervisor, Ion Androutsopoulos. His mentorship, insight, and unwavering support have been invaluable throughout the development of this thesis. I feel privileged to have worked under his guidance and to have benefited from his extensive expertise.

I am also sincerely grateful to my co-supervisor, John Pavlopoulos, for his continuous support, thoughtful feedback, and encouragement to pursue research at the intersection of Natural Language Processing and the Digital Humanities.

I am especially thankful to Yannis Assael, whose advice, generous support, and contagious enthusiasm kept me motivated even in the most challenging moments.

Special thanks go to PhD student Alessandro Locaputo for his valuable feedback during his visit, and to the AUEB NLP Group for providing a stimulating and collaborative research environment.

Finally, I am deeply grateful to my family and friends for their unwavering encouragement and patience throughout this journey.

Contents

1 Introduction

Inscriptions engraved in stone are our most direct witnesses to the daily lives, beliefs, and administrative practices of the ancient Greeks. Yet, centuries of weathering, breakage, and loss render many inscriptions fragmentary, leaving scholars to piece together scant traces of names, places, and dates. Recent advances in deep learning, most notably transformer-based architectures, have revolutionized our ability to restore and attribute these damaged texts. Models such as Pythia [ASP19] and Ithaca [Ass+22] can now propose plausible restorations, assign geographic origins, and predict engraving dates with impressive accuracy. However, their inner workings remain largely opaque: like many "black-box" neural systems, they offer predictions without exposing the reasoning behind them.

This opacity poses a critical barrier in disciplines such as epigraphy and digital humanities, where interpretability is not a luxury but a necessity. Historians and philologists require transparent, evidence-grounded explanations to validate machine-generated hypotheses against established historical knowledge. Without clear justifications, even highly accurate models risk being mistrusted or misapplied in scholarly research. Consequently, explainable AI (XAI) techniques have emerged as a vital complement to high-performing models, aiming to make their predictions intelligible and actionable for domain experts.

1.1 Motivation and Problem Statement

While Ithaca has set new standards in dating and restoring ancient Greek inscriptions, its predictions currently rely on internal attention scores and raw gradients that are difficult for historians to interpret. Early attempts at saliency mapping (e.g., multiplying gradients by input values [Shr+16]) often produce noisy heatmaps, highlighting many tokens or (even worse) characters simultaneously and offering little insight into which names or terms drove the model's decision. To bridge this gap, we must develop a systematic pipeline that (a) generates robust, theoretically grounded attributions at both the character and word levels, (b) normalizes and fuses these signals into coherent token-wise heatmaps, and (c) rigorously evaluates their quality through quantitative metrics.

Contributions

This thesis enhances the interpretability of Ithaca's date-attribution component through:

1. *Unified Explainability Pipeline.* We introduce a unified pipeline that integrates multiple post-hoc explainability methods, producing normalized, token-level saliency maps at both the character and word embedding levels without modifying the original date output.

2. *Comprehensive Quantitative Evaluation.* We introduce a suite of ranking and classification metrics to compare different explainability methods explained in Chapter 3 on a curated "low-variance" bigram-annotated dataset called the Onomastics subset (subset of the Onomastics dataset introduced by Assael et al. [Ass+22]).

Overall, this thesis addresses the following central research question: *How can the explanations produced by Ithaca's chronological-attribution component be improved?* Note that Ithaca can also predict missing characters, as well as the geographical origins of Ancient Greek inscriptions; however, this thesis focuses on its chronological attributions only.

1.2 Thesis Structure

Chapter 1: Introduction

This chapter highlights the importance of interpretability and explainability in AI-driven research tools, such as Ithaca, which epigraphers utilize to date ancient inscriptions. We formalize our central problem: how to integrate post-hoc explainability into Ithaca's date-attribution pipeline without sacrificing its state-of-the-art accuracy. Finally, we present the thesis objectives: to design a unified explainability framework and to evaluate its impact on dating performance rigorously.

Chapter 2: Background & Related Work

This chapter is a focused literature review. We first recap the evolution of deep-learning methods in digital epigraphy, highlighting their high accuracy and "black-box" nature. We then survey existing post-hoc explainability techniques, discussing their theoretical properties and prior applications in natural language processing and computer vision. Finally, we pinpoint the challenges posed by sparse, context-dependent inscription data and argue for the need to systematically compare the explainability techniques before integrating them into Ithaca.

Chapter 3: Implemented Methods

In this chapter, we detail the suite of post-hoc explainability techniques we integrated into

Ithaca's date-attribution workflow. We begin by presenting our gradient-based baseline (multiplying gradients by input signals) computed over both character- and word-level embeddings. We then introduce more advanced explainability techniques, most prominently Integrated Gradients, outlining their theoretical motivation, how we instantiate them in Ithaca (choice of baselines, step schedules), and how we collapse their high-dimensional attributions into token-wise heatmaps. Finally, we implement additional explainer families that slot into the same extraction-harmonization-visualization framework, setting the stage for their quantitative evaluation in Chapter 5.

Chapter 4: Data

In this chapter, we leverage the Onomastics benchmark—a curated subset of the *Packard Humanities Institute* (PHI) Greek Inscriptions corpus (I.PHI)[1], designed for name-based chronological attribution of ancient Greek inscriptions, first introduced by Assael et al. [Ass+22]. We detail its composition, including LGPN[2]-derived name bigrams and their date distributions, and explain how we extracted a focused "low-variance" bigram testbed for explainability experiments. Finally, we present an exploratory analysis of bigram frequencies, the distribution of bigrams per inscription, and co-occurrence patterns to illuminate the dataset's key characteristics and motivate our explainability studies.

Chapter 5: Experiments and Results

This chapter details our empirical evaluation of the explainability methods introduced in Chapter 3. We begin by describing the experimental setup, including our study of aggregation schemes (sum, max, avg) and granularity levels (character, word, combined), using retrieval-style metrics (MRR, MAP, nDCG@2) on the low-variance bigram testbed of Chapter 4 to select a default saliency aggregation. We then present the main system evaluation, which includes quantitative assessments of each explainability technique using ranking (MRR, MAP, nDCG) and classification metrics (Precision, Recall, F1, AUC).

Chapter 6: Conclusions and Future Work

This closing chapter reflects on the principal achievements of our work. We synthesize how the results demonstrate that meaningful, token-level saliency maps can be produced without undermining Ithaca's dating accuracy, and we highlight the practical guidelines derived for presenting attributions in scholarly workflows. We then discuss the limitations of our current work and the need for more diverse testbeds and propose avenues for future work. These include extending our framework to other attribution tasks, exploring additional explainability methods and aggregation strategies, and assembling richer, expert-annotated evaluation datasets to validate and refine model explanations further.

[1] `https://inscriptions.packhum.org/`
[2] `https://www.lgpn.ox.ac.uk/`

2 Background and Related Work

The analysis and interpretation of historical texts have long relied on expert intuition, fragmented evidence, and comparative research. Over the years, advancements in machine learning, primarily through the development of transformer-based models, have introduced new tools for studying the past. These models offer powerful capabilities of restoring damaged inscriptions, predicting their geographical origin, and estimating their engraved date [Ass+22]. However, despite their accuracy, such systems often function as "black boxes", producing results without clear justification [Li+22] [VEA22].

In fields of study like epigraphy, where interpretability needs to be transparent, the lack of explainability is a pressing issue that limits the usefulness of artificial intelligence in research [Gat25]. For the reliability and integration of these tools into the academic workflow, it's crucial to understand what predictions they make and why they make them [Li+22].

2.1 Background

Over the years, the use of artificial intelligence in historical research has experienced substantial progress in machine learning technologies. The field has evolved from early rule-based methods [SWM17] to today's sophisticated deep learning architectures, which enable the analysis of increasingly complex historical sources with greater precision. Recent advances in interpretability further help historians understand and trust these models' outputs [Mün+24].

Deep learning, and in particular transformer architectures, now underpin most modern NLP pipelines. For readers seeking an overview of these foundational concepts, several comprehensive surveys are available. Taye et al. [Tay23] provide a general introduction to deep learning fundamentals, while Lin et al. [Lin+22] survey the evolution and mechanisms of transformer-based models. For a broader understanding of explainability in AI, Hassija et al. [Has+24] and Molnar [Mol25] offer detailed reviews of intrinsic interpretability methods, where the model is inherently transparent (e.g., decision trees or linear models), and post-hoc interpretability methods, where external techniques such as saliency maps, SHAP, or Integrated Gradients are used to explain the predictions of otherwise opaque models. Additionally, recent studies by Li et al. [Li+23] and Silva et

al. [SSN24] focus specifically on attribution techniques for large language models, making them particularly relevant for the interpretability analysis presented in this thesis.

2.1.1 Deep Learning for Ancient Text Restoration and Attribution

Recent advancements in machine learning have begun to tackle challenges in digital epigraphy and the study of ancient inscriptions.

Pythia: The First Deep Learning Model for Restoring Ancient Greek Inscriptions
A landmark effort was Pythia, a sequence-to-sequence RNN model that could propose restorations for damaged ancient Greek texts [ASP19]. Trained on the *Packard Humanities Institute* (PHI) Greek Inscriptions corpus (I.PHI)[1]—a large digital collection of Greek epigraphic texts widely used in digital classics—Pythia achieved a 30% character error rate, significantly outperforming human epigraphists who had a 57% error rate. Moreover, in 75% of cases, the correct restoration was among Pythia's top 20 hypotheses, highlighting its potential as an assistive tool in digital epigraphy [Som+23].

Ithaca: Transformer-Based Model for Restoration and Attribution
Building on on the success of Pythia, Ithaca is a transformer-based architecture designed to restore missing text, identify an inscription's origin, and predict its endgraving date [Ass+22]. Ithaca was also trained on the PHI dataset, and achieved state-of-the-art results on these tasks. Ithaca's design emphasized collaboration with historians and built-in interpretability features.

Ithaca demonstrated a substantial improvement over previous approaches to the restoration and analysis of ancient inscriptions. On the task of automatic text restoration, Ithaca achieved an accuracy of 62% in reconstructing damaged Greek inscriptions on a held-out test set. Notably, when used in an interactive setting—where historians could review and select among Ithaca's top restoration hypotheses—the success rate of expert restorers increased dramatically, from 25% (working unaided) to 72% (with Ithaca's assistance), highlighting the potential for effective human–AI collaboration. Beyond restoration, Ithaca was also able to attribute the geographic provenance of inscriptions with 71% accuracy and predict their date within an average margin of 30 years from the ground-truth date. These results indicate that Ithaca not only provides strong predictions on its own, but also significantly enhances the capabilities of human experts when incorporated into the research workflow. A key factor in its adoption is its collaborative and interpretable design: instead of producing a single opaque output, Ithaca presents multiple ranked hypotheses with associated confidence scores, enabling historians to critically evaluate and contextu-

[1] `https://inscriptions.packhum.org/`

alize the model's suggestions [Ass+22]. We return to Ithaca in more detail in Section 2.1.2, where we discuss its architecture and relevance for this thesis.

Advancements Beyond Pythia and Ithaca

Several studies have explored deep learning techniques for the restoration and attribution of ancient texts [Som+23]. RNNs have shown strong performance for Akkadian [Fet+20] and Linear B [PKO23]. Transformer-based models have further advanced the field by introducing a flexible blank-filling architecture [She+20] and a BERT for Latin text restoration [BB20]. Lazar et al. (2021) achieved 83% accuracy in restoring Akkadian cuneiform [Laz+21], and Kang et al. (2021) reported 89% top-10 accuracy on Korean historical records [Kan+21].

Borkar and Smith (2024) [BS24] used transformer-based OCR to restore damaged texts with notable success, and Wang et al. (2023) developed GujiBERT and GujiGPT for ancient Chinese texts, demonstrating strong performance on multiple NLP tasks [Wan+23]. These works highlight the potential of transformer models to generalize across scripts and enhance collaboration between AI and human experts.

Aeneas: Multimodal Contextualization

Building on the success of Ithaca, recent developments have introduced Aeneas, a multimodal generative neural network designed for contextualizing ancient Latin inscriptions [Ass+25]. Aeneas represents a significant advancement beyond previous models by introducing three key innovations that address limitations in digital epigraphy.

First, Aeneas combines textual input with visual data, processing both inscription transcriptions and associated images when available. The model employs a transformer-based decoder enhanced with a shallow visual neural network for image processing, proving particularly valuable for geographical attribution tasks where material, style, and layout cues are crucial for historical reasoning [Ass+25].

Second, Aeneas introduces an advanced contextualization mechanism to assist historians by retrieving relevant ancient Latin inscriptions from the Latin Epigraphic Dataset (LED), which is a unified and machine-actionable collection of Latin inscriptions, created by combining three major databases: the Epi graphic Database Roma (EDR)[2], the Epigraphic Database Heidelberg (EDH)[3] and the Epigraphik-Datenbank Clauss-Slaby ETL (EDCS_ETL)[4]. Instead of searching for exact word-for-word matches, Aeneas identifies "parallel" inscriptions—texts that are similar in meaning, cultural background, social function, or historical context. These parallels are not translations but inscriptions that share linguistic patterns, formulas, or provenance with the inscription being studied. Aeneas

[2] `https://www.edr-edr.it`
[3] `https://edh.ub.uni-heidelberg.de`
[4] `https://github.com/sdam-au/EDCS_ETL`

creates "historical fingerprints" for each text, capturing semantic and functional relationships that go beyond literal text comparison. By finding these meaningful parallels, historians gain important insights to help interpret fragmentary or damaged inscriptions, date and locate them more accurately, and build a richer understanding of their historical significance. This retrieval process dramatically speeds up research by providing contextually relevant examples that would otherwise take experts much longer to uncover manually.

Third, Aeneas pioneers the capability to restore texts of arbitrary length, breaking through previous limitations that required knowing the exact length of missing segments [Ass+25]. This breakthrough enables more flexible restoration of fragmentary inscriptions where the extent of damage is unknown.

Expert evaluation involving 23 epigraphers demonstrated Aeneas's practical value: the system achieved dating accuracy within 13 years compared to 31 years for human experts working alone, while collaborative human-AI workflows improved expert confidence in key interpretive tasks by 44% [Ass+25]. The contextualization feature proved particularly transformative, reducing the time for identifying relevant parallels from days to minutes while maintaining scholarly rigor.

2.1.2 Ithaca: A Transformer for Ancient Greek Inscriptions

Of particular interest in this thesis is Ithaca's ability to predict the date of inscriptions, which involves classifying a given text into date intervals. Recent work demonstrated Ithaca's ability to predict the dates, with the average date prediction being within 28.7 years of the ground-truth date interval [Ass+22]. From the architectural perspective, Ithaca incorporates character-level and word-level embeddings, which are combined and processed through eight stacked Transformer Layers. A more detailed explanation of how the character- and word-level embeddings are combined, together with an illustration of Ithaca's architecture, is provided in Section 3.1. The model consisted of three task-specific output heads: one for restoration, one for geographical attribution, and one for date attribution. Notably, Ithaca outputs a probability distribution for chronological attribution over discrete 10-year intervals from 800 BCE to 800 CE rather than predicting a single year. This probabilistic approach reflects historical uncertainty and offers more interpretable insights [Ass+22].

Ithaca produces saliency maps [SVZ13] for its interpretability via gradient-based attribution techniques, especially multiplying gradients by input signals [Shr+16], which will be analyzed in Section 3.1. Saliency maps are visual representations that highlight which parts of an input most strongly influence a model's prediction. In the context of ancient inscriptions, a saliency map shows which words or characters the model considers most

important for dating an inscription, typically displayed as color-coded overlays where brighter colors indicate higher importance. A token is considered "salient" when it has high influence on the model's decision-making process, meaning that making a slight modification to the token's input embedding—the numerical vector that represents it in the neural network—would significantly affect the model's output prediction. These maps help historians understand which parts of an inscription the model found most informative for its decisions. More about how saliency maps work will be discussed in Section 3.1. For example, when predicting the date of an Athenian decree, Ithaca focused attention on the name "Νικίας" and the word "στρατεγοίς" both historically anchored to the 5th century BCE. Such outputs support historians by surfacing clues they can critically evaluate. Ithaca was also evaluated in a human-in-the-loop setting, where historians improved their restoration accuracy by leveraging model suggestions. This collaboration demonstrates Ithaca's operation not as a replacement for historians but as a powerful assistive tool grounded in domain-specific understanding [Ass+22].

However, despite these promising results, significant challenges in making transformer-based models like Ithaca truly transparent to domain experts remain. While current saliency maps provide some insight into model behavior, they often produce noisy visualizations that highlight many tokens (characters or words) simultaneously, making it difficult for historians to identify the specific evidence driving chronological decisions. This interpretability gap—between what the model can predict and what historians can understand about those predictions—represents a fundamental barrier to widespread adoption of AI tools in digital epigraphy. Addressing these explainability challenges through systematic comparison and enhancement of attribution methods forms the central motivation of this research.

2.2 Explainability

As deep-learning systems grow in complexity and predictive power, their inner functionality and interpretability have become increasingly important, especially in fields of study that depend on transparency, such as DH (Digital Humanities). Explainable AI (XAI) within those fields has been highlighted as essential for gaining trustful insights into historical processes [Dob21]. Explainability, especially in post-hoc methods, involves generating human-understandable justification for decisions after predictions are made [VEA22].

2.2.1 Interpreting Transformer Models in NLP Applications

Transformers added a revolutionary touch to NLP by enabling models to understand context through self-attention mechanisms. Unlike traditional recurrent neural networks,

which process input sequentially and often struggle with long-range dependencies, transformers can analyze entire sequences simultaneously [Vas+17]. This structural advantage allows for a more nuanced and globally informed language representation [RKR21]. Pretrained models such as BERT and the GPT models have demonstrated exceptional performance across various NLP tasks, establishing transformers as the foundation of modern natural language processing [Dev+19] [Bro+20]. Transformer models have also been adapted for domain-specific applications, including the including the restoration, chronological classification, and geographical attribution of historical texts [Ass+22].

Despite the success of transformer models like Ithaca in prediction tasks, a persistent challenge is making their decisions transparent. Transformer architectures rely on complex attention mechanisms and high-dimensional representations, which humans do not understand [RKR21].

2.2.2 Post-Hoc Explanation Techniques for Model Interpretability

To address this challenge of transparency and interpretability in transformer models, researchers have developed post-hoc explainability techniques that attribute model predictions to input features, helping to bridge the gap between complex model reasoning and human understanding [DK17] [Bar+20]. Feature selection is generally used before or during the model training while feature attribution is used to explain an already trained model (post-hoc explanation). Feature attribution, includes techniques such as gradient-based methods [Sel+19], perturbation-based methods (modifying inputs and observing output changes) [ZF13], and surrogate models (approximating complex models with interpretable ones such as decision trees) [RSG16]. These methods aim to explain a model's prediction by highlighting the relevance of each input feature [Mol25].

2.2.3 Attribution Methods in Practice: SHAP, LIME, IG and LRP

Integrated Gradients (IG) is an attribution method introduced by Sundararajan et al. (2017) [STY17]. It assigns an importance score to each input feature by integrating the model's gradients as the input transitions from a baseline to its actual value. Intuitively, IG accumulates how much each character or token influences the prediction as we transition from a neutral input (e.g., a blank inscription) to the real inscription. This method satisfies certain desirable axioms (sensitivity and implementation invariance) that many earlier methods lacked [STY17]. IG has been widely applied to interpret deep models in vision and NLP, as it only requires access to the model's gradient and does not alter the model's internal parameters. In the context of transformer-based language models, IG

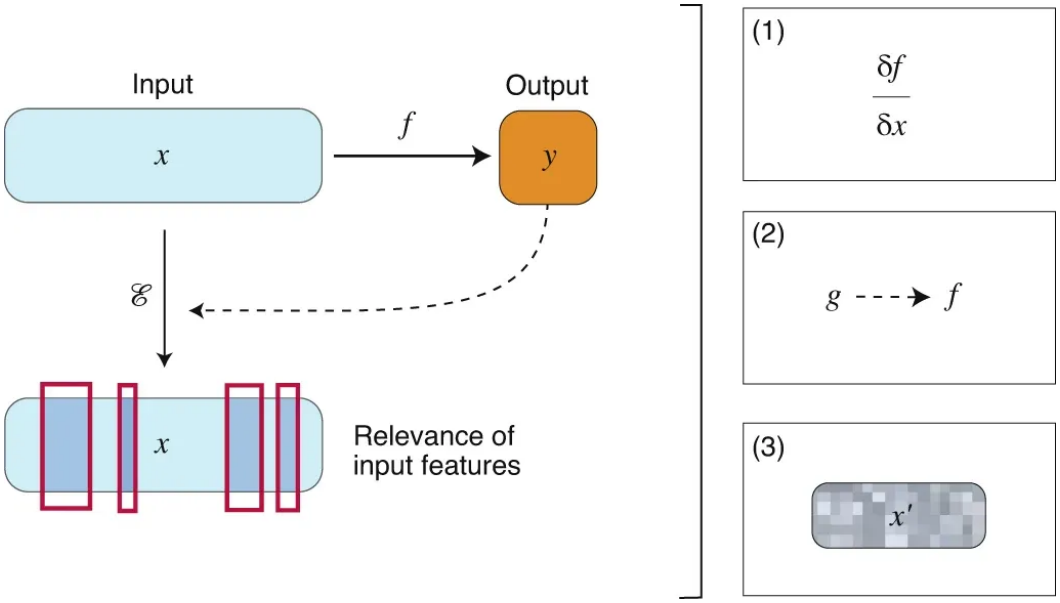


Fig. 2.1: (1) Gradient-based methods, (2) Surrogate methods, (3) Perturbation-based methods. Figure taken from [JGS20]

can highlight which words or letters were most responsible for a classification [RKR21]; for instance, identifying which parts of an ancient text led Ithaca to predict a specific date. Another common approach, Layer-Wise Relevance Propagation (LRP), back-propagates a prediction score through the network layers to distribute relevance scores to the input features [Bac+15] [Mon+19]. LRP has been used to explain decisions in domains from image recognition to document analysis [Bac+15] [Arr+16]. IG and LRP produce visual explanations that help researchers see what the models considers essential [Sam+21].

2.2.4 Visual Explainability in AI

As deep learning models advance and become more complex, transparency in decision-making has become crucial. Visual explainability typically acts as the presentation layer of attribution methods, transforming token-level importance scores into visualizations such as saliency maps. This is essential in academic fields like epigraphy, where the reliability of the model's interpretation depends on the scholars' ability to understand, annotate, and validate the model's reasoning [Ass+22].

Early visualization techniques such as saliency maps—often derived from raw gradient computations—have offered limited interpretive value in text-based models. Because they use only local gradient information, saliency maps are easily affected by noise and non-linear effects in the model. As a result, they often highlight too many tokens at once, making it hard to see which ones really mattered, especially in sequential tasks like language modeling [Arr+16]. More advanced interpretability methods, like Integrated Gradients (IG), provide more reliable insights [STY17]. These approaches help contextualize

model behavior, allowing researchers to understand which words or characters most influenced predictions, such as Ithaca's chronological attribution, thereby increasing both interpretability and scholarly trust within digital epigraphy [Ass+22].

2.2.5 Visual Explanations in Ithaca

Regarding Ithaca, such attribution techniques were the key to its interpretability-focused design. The model produces saliency maps as a visual aid for restoration and attribution tasks. These maps highlight which characters or words in the input inscription influence the model's decision most. To further illustrate this point, Figures 2.2 and 2.3 present additional examples of Ithaca's token-level saliency maps. In Figure 2.2, the bigram 'πυργος μιχαηλ' is sharply localized, showing how attribution methods can align with historically meaningful markers. Similarly, in Figure 2.3, the personal name 'τιβεριος κλαυδιος' and the demotic 'σινωπευς' are emphasized, both of which are informative for chronological attribution. These examples demonstrate how saliency maps can highlight contextually significant tokens, thereby enhancing transparency in the model's reasoning. At the same time, they reveal that attribution quality can vary, underscoring the importance of examining multiple cases.

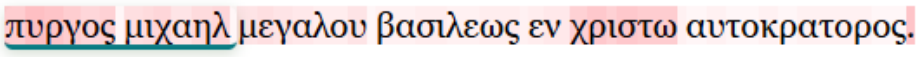


Fig. 2.2: Ithaca's example saliency map for chronological attribution. The saliency overlay highlights the bigram 'πυργος μιχαηλ' as having the highest influence on the model's dating prediction. This alignment between the model's highlighted features and historically meaningful markers provides a transparent justification for the output.

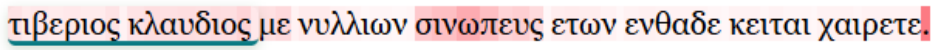


Fig. 2.3: Ithaca's example saliency map for chronological attribution. The saliency overlay highlights the personal name 'τιβεριος κλαυδιος' and the demotic 'σινωπευς' as influential for the model's dating prediction. These features correspond to historically meaningful markers, providing a transparent rationale for the model's output.

Similarly, for the restoration task, Ithaca does not just output one guess for a missing fragment but offers a ranked list of the top 20 suggestions with probabilities. This allows the researchers to consider multiple plausible restorations side by side, improving the interpretability of the system's output by presenting alternative hypotheses. Instead of outputting just one year, Ithaca generates a probability distribution over date intervals for the dating task [Ass+22].

3 Implemented Methods

3.1 Baseline Interpretability: Ithaca's Gradient ⊙ Input Saliency Maps

To provide interpretability, Ithaca leverages the Gradient ⊙ Input method to generate saliency maps that highlight the relative importance of each character and word in the input sequence for the model's prediction [Ass+22] [SVZ13]. As shown in Figure 3.1, each inscription is encoded at both the character and word level, concatenated with positional embeddings, and then processed through stacked transformer layers. The model outputs predictions for text restoration, geographical attribution, and chronological attribution, while Gradient ⊙ Input saliency maps are computed with respect to the final embedding representations. This design allows interpretability analyses to focus directly on how the model distributes importance across tokens, which is critical for evaluating its behavior in chronological attribution tasks.

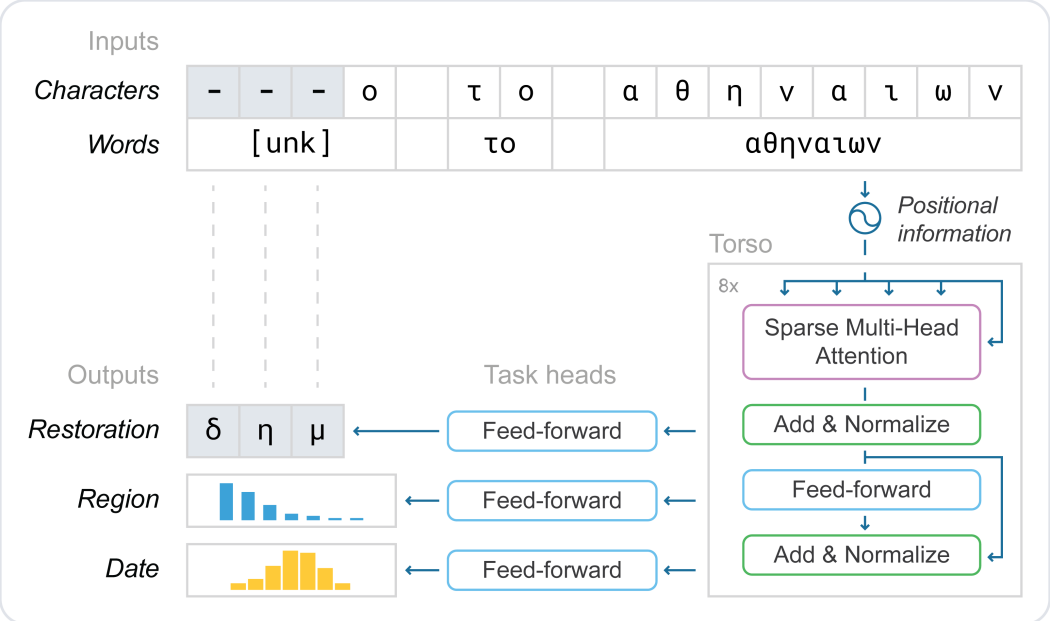


Fig. 3.1: Overview of the Ithaca architecture. Each input inscription is represented at the character and word level and processed through stacked transformer layers with positional information. The model outputs predictions for text restoration, geographical region, and chronological attribution. Gradient ⊙ Input saliency maps are computed using the final embedding representations and the output layer for each task. Figure taken from [Ass+22].

Feature Extraction and Representation

Ithaca first encodes each inscription in two parallel ways: at the character level and at the word level. This produces two separate sequences of embeddings for each position in the input. These embeddings are concatenated and augmented with positional information to capture the meaning of the inscription. The stacked transformer blocks then produce a context-aware representation for each word token, where each representation combines both the word-level embedding and its corresponding character-level embedding. These token representations are then used for downstream prediction and attribution [Ass+22].

Gradient $\odot$ Input Explanations

To quantify the influence of each input token on the model's output, the Gradient $\odot$ Input method is applied [SVZ13]. Specifically, the importance score for the i-th token is calculated as the element-wise product of the input embedding and the gradient of the output with respect to that embedding. The gradient alone shows how sensitive the output is to changes in the token, but it ignores the token's actual value. Gradient $\odot$ Input combines both, capturing not just potential sensitivity but also the token's real contribution to the prediction.

The saliency score for token i is computed as:

$$\text{Saliency}_i = \left\| \frac{\partial F(\mathbf{x})}{\partial \mathbf{x}_i} \odot \mathbf{x}_i \right\|_2 \tag{3.1}$$

where $\mathbf{x}_i$ is the embedding of token i and $\frac{\partial F}{\partial x_i}$ is the gradient of the output score with respect to x_i.

The L2 norm aggregates per-dimension attribution scores into a single scalar for each token, facilitating visualization and comparison [Arr+16].

In the context of chronological attribution, $F(\mathbf{x})$ refers to Ithaca's predicted date interval as the most likely date for the inscription. The saliency map is therefore computed by taking the gradient of this scalar output with respect to the input embeddings, following standard practices in neural model interpretability. This results in a scalar saliency value per token, reflecting how much a small change in the embedding of that token would affect the model's output.

This method can be seen as a first-order approximation of the model's output sensitivity to the input token embedding, grounded in the Taylor expansion of the model's output function $F(\mathbf{x})$ [STY17].

Formally, the first-order Taylor expansion of F around a baseline input $\mathbf{x}_0$ is given by:

$$F(\mathbf{x}) \approx F(\mathbf{x}_0) + \nabla F(\mathbf{x}) \cdot (\mathbf{x} - \mathbf{x}_0)$$

If we take $\mathbf{x}_0 = \mathbf{0}$ (i.e., a zero baseline), then this corresponds to replacing each token embedding with the all-zero vector:

$$F(\mathbf{x}) \approx F(\mathbf{0}) + \nabla F(\mathbf{x}) \cdot \mathbf{x}$$

The element-wise product $\nabla F(\mathbf{x}) \odot \mathbf{x}$ thus comes from the first-order Taylor expansion, which says we can approximate a complicated model locally as a linear function of its inputs. In this view, each input dimension makes a weighted contribution to the output, where the weight is given by the gradient. This is different from using the gradient alone, which only shows how the output would change if we nudged the input, and different from using the input alone, which only shows how strongly a token is represented. By combining them, Gradient $\odot$ Input captures how much each token is actually contributing to the current prediction. One can either sum these contributions or aggregate them via a norm. In this thesis, we use the L2 norm (Eq. 3.1), which yields non-negative scalar scores by squaring all contributions.

Implementation in Ithaca

In Ithaca, the model receives input at two linguistic levels:

- *Character-level input:* $x^{\text{char}} \in \mathbb{R}^{L \times D}$, where L is the sequence length (number of characters) and D is the embedding dimension (the size of the vector used to represent each character).

- *Word-level input:* $x^{\text{word}} \in \mathbb{R}^{L \times D}$. To align with the character sequence, each word embedding is repeated for all characters belonging to that word, so that x^{char} and x^{word} have the same length L. Both character and word embeddings share the same dimensionality D, allowing them to be concatenated into a $2D$-dimensional representation per position.

To generate attribution maps for a specific prediction task (e.g., date), Ithaca computes the gradient of the logit output for the predicted class with respect to both x^{char} and x^{word}. These gradients are then multiplied element-wise with the embeddings of the respective input tokens and projected to a scalar score for each position in the input sequence (i.e., each character, augmented with its word embedding).

Let:

$$g_i^{\text{char}} = \frac{\partial F}{\partial x_i^{\text{char}}}, \qquad g_i^{\text{word}} = \frac{\partial F}{\partial x_i^{\text{word}}}$$

Then:

$$\text{Saliency}_i^{\text{char}} = \left\| g_i^{\text{char}} \odot x_i^{\text{char}} \right\|_2, \qquad \text{Saliency}_i^{\text{word}} = \left\| g_i^{\text{word}} \odot x_i^{\text{word}} \right\|_2$$

Finally, the total saliency map is formed by summing both contributions:

$$\text{Saliency}_i = \texttt{clip}\left(\text{Saliency}_i^{\text{char}} + \text{Saliency}_i^{\text{word}},\, 0,\, 1\right) \tag{3.2}$$

where `clip` denotes element-wise clipping of the saliency values to the $[0, 1]$ interval, i.e., values below 0 are set to 0 and values above 1 are set to 1. Here, the index i always refers to the i-th *character position* in the sequence. At each position, the model concatenates the embedding of that character with the embedding of the word to which the character belongs, so both x_i^{char} and x_i^{word} (and their gradients) are defined for every character index.

These saliency scores can be visualized as heatmaps over the input, highlighting the regions most influential in the model's dating decision, as illustrated in Figure 3.2.

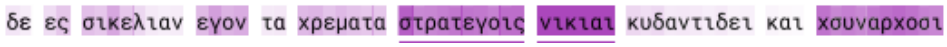


Fig. 3.2: Example Gradient ⊙ Input saliency map for chronological attribution. The saliency overlay highlights the words "στρατεγοις" and "νικιαι" as having the highest influence on the model's dating prediction ("Athens, 414/3 BC"). Such explanations align with historical reasoning and provide transparent justification for the model's output.

3.2 Integrated Gradients

While the Gradient ⊙ Input method provides a simple and efficient baseline for model interpretability, it is sensitive to noise and non-linearities. Raw gradient methods can be misleading—for example, if a sigmoid unit is saturated (very close to 0 or 1), the gradient is nearly zero and hides the importance of a token, while near a sharp boundary the gradient can spike and exaggerate a single token's role. As a result, raw gradient values can either vanish in saturated regions or spike unpredictably, producing attributions that do not faithfully reflect feature importance. To overcome these limitations, we adopt and systematically evaluate the Integrated Gradients (IG) method [STY17] for chronological attribution with Ithaca. In this thesis, we focus on the task of chronological attribution, as accurate dating of inscriptions is one of the most critical problems for historians and epigraphers, and it provides a clear setting for evaluating interpretability. Other prediction

tasks supported by Ithaca, such as restoration and geographical attribution, are beyond the scope of this study and are left for future work.

The theoretical motivation for Integrated Gradients stems from a limitation of raw gradient-based attribution: the local gradient at the input may be close to zero if the model output is nearly constant in that region (e.g. a sigmoid in a saturated state). In such cases, simple saliency maps would suggest that the feature is unimportant, even though the feature may have been crucial in moving the output from the baseline $\mathbf{x}'$ to the final prediction at $\mathbf{x}$.

The core idea of IG is to attribute the prediction to each input by accumulating gradients along a continuous path from a reference baseline $\mathbf{x}'$ to the input $\mathbf{x}$. In doing so, IG captures not only the local sensitivity at the final input, but the entire change in the model output as the input is gradually introduced starting from the baseline. The path from baseline to input need not be sensitive everywhere, but by integrating the gradients along it, IG guarantees that the total attribution matches the change in the model's output between baseline and input.

$$\mathrm{IG}_i(\mathbf{x}) = (x_i - x_i') \odot \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_i} \, d\alpha \tag{3.3}$$

Where:

- $\mathbf{x}$ is the model's input embedding sequence, i.e., a matrix of shape $[L, D]$, where L is the input sequence length (number of words or characters), and D is the embedding dimension. As discussed in Section 3.1, both the character-level embeddings and the word-level embeddings are projected to the same dimensionality D. To ensure proper alignment and concatenation, the sequence length L is identical for both character and word embeddings, with each word embedding repeated across the characters that constitute the word.

- x_i is the embedding vector for the i-th token ($x_i \in \mathbb{R}^D$).

- $\mathbf{x}'$ is a baseline embedding sequence (same shape as $\mathbf{x}$), representing a "neutral" or reference input. In practice, this can be either a sequence containing L copies of the all-zeros embedding (canonical baseline) or L copies of the centroid (mean) of the embedding matrix.

- x_i' is the baseline embedding for the i-th token.

- $\alpha \in [0, 1]$ is the interpolation parameter along the path from baseline to input.

- $F(\cdot)$ is the scalar model output with respect to the prediction of interest. For chronological attribution in Ithaca, $F(\cdot)$ is the logit of the predicted class for the given inscription.

- $\frac{\partial F}{\partial x_i}$ is the gradient of the model output with respect to the embedding of token i.

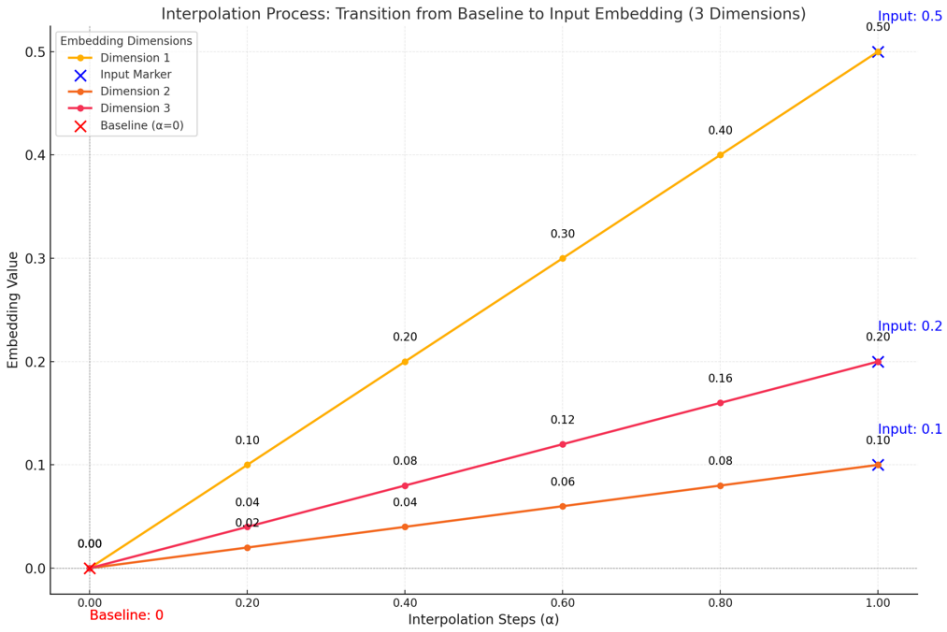


Fig. 3.3: Visualization of the Integrated Gradients interpolation process for a single token embedding in three dimensions. Each line traces the value of one embedding dimension as the interpolation parameter α transitions from the baseline (often zeros) to the actual input embedding. IG computes the gradient of the model output at each interpolated step, which are then integrated to form the final attribution.

We now illustrate the Integrated Gradients (IG) method using a sequence of visualizations that build intuition for how the attribution is computed:

- **Fig. 3.3** illustrates the interpolation process in a low-dimensional embedding space. Each embedding dimension increases linearly from an all-zeroes baseline to the input vector as α varies from 0 to 1. This stepwise path highlights how IG samples intermediate embeddings between the baseline and the input.

- **Fig. 3.4** offers an alternative illustration of IG: gradients are sampled at interpolated points along a straight-line path over the model's output surface. These gradients are integrated to yield an attribution vector per input feature.

The integral is approximated by a Riemann sum over m discrete steps:

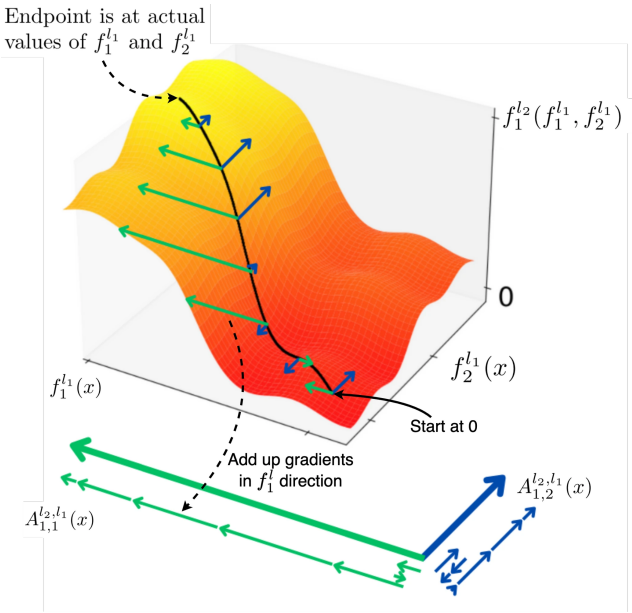


Fig. 3.4: Schematic illustration of the Integrated Gradients method. Attributions are computed by integrating the gradient of the model output along a straight path from a baseline input to the actual input, accumulating the contribution for each input feature. Figure taken from [Shi22].

$$\mathrm{IG}_i(\mathbf{x}) \approx (x_i - x_i') \odot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F\left(\mathbf{x}' + \frac{k}{m}(\mathbf{x} - \mathbf{x}')\right)}{\partial x_i} \tag{3.4}$$

where m is a hyperparameter controlling the granularity of the integration (typically $m = 50$).

Aggregation and Visualization

For each character position i, Integrated Gradients produces two attribution vectors: one for the character embedding and one for the aligned word embedding. Each is reduced to a scalar score using the L2 norm:

$$\mathrm{Saliency}_i^{\mathrm{char,IG}} = \left\|\mathrm{IG}_i^{\mathrm{char}}(\mathbf{x})\right\|_2, \qquad \mathrm{Saliency}_i^{\mathrm{word,IG}} = \left\|\mathrm{IG}_i^{\mathrm{word}}(\mathbf{x})\right\|_2 \tag{3.5}$$

The total saliency at position i is then obtained by summing the two contributions and normalizing via clipping:

$$\mathrm{Saliency}_i^{\mathrm{IG}} = \texttt{clip}\left(\mathrm{Saliency}_i^{\mathrm{char,IG}} + \mathrm{Saliency}_i^{\mathrm{word,IG}},\, 0,\, 1\right) \tag{3.6}$$

where `clip` denotes element-wise clipping of the saliency values to the $[0, 1]$ interval, i.e., values below 0 are set to 0 and values above 1 are set to 1. This ensures comparability with Gradient $\odot$ Input and supports intuitive visualizations.

Integrated Gradients (IG) admits a number of implementation choices that can affect attribution quality. In our Ithaca experiments we systematically evaluated four IG variants along two axes: choice of baseline ($\mathbf{x}'$), as well as number and selection of integration steps (m). In all cases, vector attributions are reduced to scalar token saliency scores using the L2 norm (Eq. 3.5), with character- and word-level contributions summed and normalized via the `clip` operation (Eq. 3.6) to the $[0, 1]$ interval for visualization. The specific IG variants are reported in the following subsections.

3.2.1 Multi-step IG with Zero Baseline

The baseline is set to the all-zeros embedding:

$$\mathbf{x}' = \mathbf{0} \in \mathbb{R}^{L \times D}$$

With $m = 50$ evenly spaced interpolation coefficients $\alpha_k = \frac{k}{m}$, $k = 1, \ldots, m$, the intermediate inputs simplify to:

$$\mathbf{x}^{(k)} = \alpha_k \mathbf{x}$$

Gradients are computed at each step and averaged:

$$\mathrm{IG}_i(\mathbf{x}) \approx x_i \odot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F(\mathbf{x}^{(k)})}{\partial x_i}$$

3.2.2 Single-step IG with Zero Baseline

Setting $m = 1$ with $\alpha = 1$ yields:

$$\mathbf{x}^{(1)} = \mathbf{x}, \quad \mathbf{x}' = \mathbf{0}$$

This simplifies to:

$$\mathrm{IG}_i(\mathbf{x}) = x_i \odot \frac{\partial F(\mathbf{x})}{\partial x_i}$$

This corresponds exactly to Gradient $\odot$ Input and is used as a baseline sanity check.

3.2.3 Multi-step IG with Input Centroid Baseline

Here, the baseline is the centroid of the current input:

$$\mathbf{x}' = \frac{1}{L}\sum_{i=1}^{L} x_i$$

This locally grounded baseline captures the average meaning of the inscription.

We use $m = 50$ evenly spaced interpolation coefficients $\alpha_k = \frac{k}{m},\ k = 1, \ldots, m$, consistent with the zero-baseline variant:

$$\mathbf{x}^{(k)} = \mathbf{x}' + \alpha_k(\mathbf{x} - \mathbf{x}')$$

This variant combines path smoothing with contextual awareness from the input's own semantics.

3.2.4 Single-step IG with Input Centroid Baseline

This variant applies Integrated Gradients with the local centroid baseline, but approximates the integral with a single step. We evaluate the gradient at the actual input ($\alpha = 1$), to ensure consistency with the single-step zero-baseline case (see Section 3.2.2):

$$\mathbf{x}^{(1)} = \mathbf{x}, \qquad \mathrm{IG}_i(\mathbf{x}) = (x_i - x_i') \odot \frac{\partial F(\mathbf{x})}{\partial x_i}$$

where x_i' is the centroid baseline embedding for token i.

This formulation differs from the zero-baseline variant only in the choice of baseline, allowing a fairer comparison across baselines.

Though rarely used in practice, this fast approximation captures deviations from the inscription's semantic average and may reduce global bias.

Summary and Evaluation Plan

All IG variants follow a harmonized processing pipeline: compute vector attributions using Eq. (3.4), reduce them to scalar saliency scores via Eq. (3.5), and aggregate character-

and word-level saliency (normalized with a `clip` operation to ensure values lie in $[0, 1]$) to produce unified token-wise visualizations. In Chapter 5, we evaluate these variants quantitatively, using ranking and classification metrics such as MRR, MAP, nDCG, Precision, Recall, F1, and AUC.

3.3 Sequential Integrated Gradients (SeqIG)

Integrated Gradients (IG) assumes that meaningful attributions can be obtained by interpolating between a baseline and the full input sequence [STY17]. However, in natural language processing, such full-sequence interpolation can yield invalid or semantically incoherent intermediate states. To address this, we adopt the Sequential Integrated Gradients (SeqIG) method [Eng23], which decomposes attribution into token-wise paths, interpolating only one token at a time while keeping the rest of the input fixed.

Notation and Setup
Let the input to the model be a sequence of token embeddings $\mathbf{S} = (x_1, x_2, \ldots, x_L)$, where each $x_i \in \mathbb{R}^D$ is the embedding of the i-th token, and L is the sequence length. For each token x_i, we construct a baseline sequence $\mathbf{S}^{(i)}$ in which only the i-th token is replaced by a fixed baseline embedding x_i', and all other tokens remain unchanged:

$$\mathbf{S}^{(i)} = (x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_L)$$

In our experiments, we use the zero vector as the baseline embedding, i.e., $x_i' = \mathbf{0} \in \mathbb{R}^D$. This choice ensures consistency with prior work [STY17]. However, unlike in the standard IG variants (see Section 3.2), we did not experiment with alternative baselines such as the centroid embedding. Extending SeqIG to centroid or context-aware baselines would provide a fairer comparison across attribution methods and may yield further insights into the role of baseline selection. We leave this as an avenue for future work.

The attribution for the j-th embedding feature of token x_i is given by:

$$\mathrm{SeqIG}_{ij}(\mathbf{S}) := (x_{ij} - x_{ij}') \cdot \int_0^1 \frac{\partial F\left(\mathbf{S}^{(i)} + \alpha(\mathbf{S} - \mathbf{S}^{(i)})\right)}{\partial x_{ij}} \, d\alpha \tag{3.7}$$

where:

- $F(\cdot)$ is the scalar model output.
- $\mathbf{S}^{(i)}$ is the baseline-modified sequence for token x_i.

- $\frac{\partial F}{\partial x_{ij}}$ is the partial derivative of the output with respect to the j-th component of x_i.
- Only the i-th token is interpolated, the rest remain fixed.

Note that in contrast to the standard IG formulation (Eq. 3.3), here the attribution is computed per embedding dimension j, so the multiplication in Eq. (3.7) is scalar.

As in the standard IG variants (Sec. 3.2), we use evenly spaced interpolation coefficients $\alpha_k = \frac{k}{m}, k = 1, \ldots, m$

This integral is approximated using a Riemann sum over m discrete steps:

$$\text{SeqIG}_{ij}(\mathbf{S}) \approx (x_{ij} - x'_{ij}) \cdot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F\left(\mathbf{S}^{(i)} + \frac{k}{m}(\mathbf{S} - \mathbf{S}^{(i)})\right)}{\partial x_{ij}} \tag{3.8}$$

Aggregation and Saliency Extraction
To obtain scalar saliency scores per token, we aggregate attributions across embedding dimensions using the L_2 norm:

$$\text{SeqIG}_i(\mathbf{S}) := \left\| \left(\text{SeqIG}_{ij}(\mathbf{S})\right)_{j=1}^{D} \right\|_2$$

This aligns with Ithaca's saliency map methodology and ensures direct comparability of results. Exploring alternative aggregation functions, such as the dot product or sum, is left for future work.

This is computed separately for character- and word-level embeddings:

$$\text{Saliency}_i^{\text{char,SeqIG}} = \text{SeqIG}_i^{\text{char}}(\mathbf{S}), \qquad \text{Saliency}_i^{\text{word,SeqIG}} = \text{SeqIG}_i^{\text{word}}(\mathbf{S})$$

The two contributions are then combined with clipping, consistent with Eq. (3.6):

$$\text{Saliency}_i^{\text{SeqIG}} = \texttt{clip}\left(\text{Saliency}_i^{\text{char,SeqIG}} + \text{Saliency}_i^{\text{word,SeqIG}}, 0, 1\right)$$

Sequential Integrated Gradients preserves input validity by interpolating only one token at a time, and yields more localized and stable attributions compared to standard IG [Eng23]. In Chapter 5, we compare SeqIG against alternative attribution methods quantitatively (e.g., MRR, nDCG, AUC), evaluating its ability to produce faithful and interpretable saliency maps for ancient inscription dating.

3.4 Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) [RSG16] approximates the behavior of a complex model in the local neighborhood of a given input by fitting an interpretable surrogate model. In our setting, LIME is used to explain Ithaca's date prediction for a specific inscription by constructing a sparse linear approximation around that input.

Let $\mathbf{S} = (x_1, \ldots, x_L)$ be the embedding sequence of an inscription and $F(\mathbf{S})$ the scalar logit corresponding to the predicted date interval. The LIME procedure proceeds as follows:

1. **Perturbation:** Generate N perturbed inputs $\{z^{(k)}\}$ by randomly masking a proportion p_{mask} of the tokens in $\mathbf{S}$. We simulate masking by replacing the selected token embeddings with zero vectors of dimension D. In our experiments, we set $N = 300$ and $p_{\text{mask}} = 0.4$.

$$z_i^{(k)} = \begin{cases} \mathbf{0} & \text{with probability } p_{\text{mask}} = 0.4 \\ x_i & \text{otherwise} \end{cases}$$

   This strategy preserves the input length and is compatible with transformer-based models not trained with masked language modeling objectives. We tuned N empirically by balancing computational cost and explanation stability. Specifically, we compared $N \in \{100, 200, 300, 500\}$ and observed that explanations became stable beyond $N = 300$, while larger values increased runtime without significant gains. We therefore fixed $N = 300$ in our experiments. Similarly, we set the masking probability to $p_{\text{mask}} = 0.4$ after testing values in the range $\{0.2, 0.3, 0.4, 0.5\}$. Lower probabilities yielded perturbations that were too close to the original input, while higher probabilities removed too much signal. The choice of $p_{\text{mask}} = 0.4$ thus provided a good trade-off between local fidelity and interpretability.

   Alternative perturbation baselines, such as replacing masked tokens with the centroid of the current sequence or with the vocabulary-wide centroid, could produce more semantically coherent perturbations, but we leave this exploration for future work.

2. **Proximity weighting:** For each perturbed sequence $z^{(k)}$, compute its cosine similarity to the original input $\mathbf{S}$, and define the corresponding weight as:

$$\pi_{\mathbf{S}}(z^{(k)}) = \exp\left(-\frac{\text{Cosine}^2(\mathbf{S}, z^{(k)})}{\sigma^2}\right)$$

where $\sigma = 0.5$ controls the locality of the neighborhood. We set $\sigma = 0.5$ as a locality hyperparameter, following common practice in kernelized LIME implementations [RSG16]. This value balances locality and sample coverage, and we found it provided stable explanations in preliminary experiments. A more exhaustive tuning of σ is left for future work.

3. **Surrogate model fitting:** Fit a sparse linear model

$$g(z) = w_0 + \sum_{i=1}^{L} w_i z_i$$

to approximate the black-box model $F(\cdot)$ in the local neighborhood of $\mathbf{S}$. Here, each perturbed input $z^{(k)}$ is represented as a binary mask vector $z^{(k)} \in \{0,1\}^L$, where $z_i^{(k)} = 0$ if token x_i is masked and $z_i^{(k)} = 1$ otherwise. For each perturbation, we query the black-box model to obtain the target output $F(z^{(k)})$, which serves as the label for the surrogate. The surrogate is trained using weighted least squares regression with the proximity weights $\pi_{\mathbf{S}}(z^{(k)})$ from Step 2 as sample weights. An ℓ_1 (Lasso) penalty is applied to encourage sparsity. If more than 10 coefficients remain nonzero after fitting, we retain only the 10 with the largest absolute values and set the rest to zero, following [RSG16]. This ensures that the surrogate model is interpretable.

Saliency Extraction

To produce token-level saliency maps, we retain only the positive weights w_i, normalize them to $[0, 1]$, and assign them separately to the character- and word-level inputs:

$$\text{Saliency}_i^{\text{char,LIME}} = \frac{\max(0, w_i^{\text{char}})}{\max_j w_j^{\text{char}}}, \qquad \text{Saliency}_i^{\text{word,LIME}} = \frac{\max(0, w_i^{\text{word}})}{\max_j w_j^{\text{word}}}$$

The total saliency at position i is then obtained by summing the two contributions and applying clipping, consistent with Eq. (3.6):

$$\text{Saliency}_i^{\text{LIME}} = \texttt{clip}\Big(\text{Saliency}_i^{\text{char,LIME}} + \text{Saliency}_i^{\text{word,LIME}},\, 0,\, 1\Big) \tag{3.9}$$

This yields sparse and locally faithful attributions that highlight the tokens most influential for the model's output in the vicinity of the original input.

3.5 SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) [LL17] assign an importance value to each input token by estimating its contribution to the model output, grounded in cooperative game theory. Specifically, SHAP attributes the prediction of a model F for an input sequence $\mathbf{S} = (x_1, \ldots, x_M)$ by computing Shapley values ϕ_i, which represent the marginal contribution of each token x_i across all possible token subsets.

Formally, the Shapley value for token x_i is defined as:

$$\phi_i(F, \mathbf{S}) = \sum_{S \subseteq \{x_1, \ldots, x_M\} \setminus \{x_i\}} \frac{|S|!(M - |S| - 1)!}{M!} \Big[F(\mathbf{S}_{S \cup \{x_i\}}) - F(\mathbf{S}_S) \Big]$$

where:

- $\mathbf{S}_S$ denotes a perturbed version of the input where only the tokens in subset S are retained (others are masked).
- $F(\cdot)$ is the scalar model output.

Since computing the exact Shapley value requires evaluating all 2^M subsets, which is intractable for long sequences, we adopt *Kernel SHAP* [LL17], a tractable approximation via local surrogate modeling:

1. **Perturbation:** Generate $N = 300$ perturbed sequences $\{z^{(k)}\}$ by randomly masking subsets of tokens in $\mathbf{S}$. Masking is implemented by replacing token embeddings with zero vectors of dimension D, consistent with the binary "feature on/off" formulation of SHAP. The number of perturbations N and the masking probability p_{mask} are hyperparameters: we set $N = 300$ for computational efficiency and $p_{\text{mask}} = 0.4$ to ensure sufficient variation while retaining semantic signal. While alternative masking strategies—such as replacing tokens with centroid embeddings to maintain plausibility—could be explored, we leave this as future work.

2. **Model Evaluation:** Query the model F for each perturbed input $z^{(k)}$ to obtain the corresponding output $F(z^{(k)})$.

3. **Surrogate model fitting:** We estimate the contribution vector $\boldsymbol{\phi} \in \mathbb{R}^M$ (Shapley values for each of the M tokens) by fitting a weighted linear model:

$$\min_{\boldsymbol{\phi}} \sum_{k=1}^{N} \left[F(z^{(k)}) - \phi_0 - \sum_{i=1}^{M} \phi_i z_i^{(k)} \right]^2 \cdot w(z^{(k)}) + \lambda \|\boldsymbol{\phi}\|_1$$

where:

- ϕ_0 is the intercept term.
- $z^{(k)} \in \{0,1\}^M$ is the binary indicator vector for the k-th perturbed input, where $z_i^{(k)} = 1$ if token i is present (unmasked) and $z_i^{(k)} = 0$ if token i is absent (masked and replaced by the baseline embedding).
- $F(z^{(k)})$ is the output of the black-box model (Ithaca) when evaluated on the perturbed input corresponding to $z^{(k)}$.
- $w(z^{(k)}) = \frac{M-1}{\binom{M}{|z^{(k)}|}\,|z^{(k)}|\,(M-|z^{(k)}|)}$ is the SHAP kernel weight that prioritizes smaller subsets [LL17].
- λ is a small regularization constant (L1 penalty) to encourage sparsity.

This formulation ensures that ϕ_i approximates the Shapley value of token i, i.e., its average marginal contribution across all possible subsets of tokens. The kernel weight reflects the intuition that smaller subsets provide clearer evidence of a token's individual contribution, while larger subsets confound effects across many tokens. By emphasizing smaller subsets, the surrogate model aligns more closely with the combinatorial definition of Shapley values, where each feature's contribution is averaged over all possible subsets [LL17].

Saliency Extraction

For each position i, we compute separate SHAP-based saliency scores for the character- and word-level embeddings:

$$\text{Saliency}_i^{\text{char,SHAP}} = \frac{\max(0, \phi_i^{\text{char}})}{\max_j \phi_j^{\text{char}}}, \qquad \text{Saliency}_i^{\text{word,SHAP}} = \frac{\max(0, \phi_i^{\text{word}})}{\max_j \phi_j^{\text{word}}}$$

The two contributions are then combined as in Eq. (3.2), with clipping to ensure all values lie within the $[0, 1]$ interval:

$$\text{Saliency}_i^{\text{SHAP}} = \texttt{clip}\Big(\text{Saliency}_i^{\text{char,SHAP}} + \text{Saliency}_i^{\text{word,SHAP}},\ 0,\ 1\Big) \tag{3.10}$$

This produces an interpretable heatmap indicating which tokens most contributed to the model's decision in a locally faithful manner.

3.6 Layer-Wise Relevance Propagation (LRP)

Layer-Wise Relevance Propagation (LRP) [Bac+15] is a decomposition-based interpretability method that attributes the model's prediction back to the input by redistributing the output relevance layer by layer. For a model output $F(\mathbf{S})$, LRP assigns each input token a relevance score R_i indicating its contribution to the prediction.

Relevance Redistribution Rule

In standard feed-forward layers, LRP redistributes the relevance score R_j of neuron j to its input neurons i using the z-rule:

$$R_i = \sum_j \frac{a_i w_{ij}}{\sum_{i' \in \text{inputs}(j)} a_{i'} w_{i'j} + \varepsilon} R_j,$$

where a_i is the activation of neuron i, w_{ij} the connection weight, and ε a stabilizer to avoid numerical instability.

Attention-Aware Layer-wise Relevance Propagation (AttnLRP)

For transformer-based architectures like Ithaca, we adopt the *Attention-Aware Layer-wise Relevance Propagation*[1] (AttnLRP) method [Ach+24], which generalizes the above rule to the structure of multi-head attention. In this setting, the weights w_{ij} are replaced by normalized attention coefficients, allowing relevance to flow between tokens in proportion to their learned attention scores. The standard z-rule is still applied in feed-forward sublayers. This extension faithfully and holistically attributes both inputs and latent representations, while maintaining computational efficiency comparable to a single backward pass.

Saliency Extraction

At the input layer, relevance scores are obtained separately for character- and word-level embeddings. Let R_i^{char} and R_i^{word} denote the relevance assigned to the character and word embedding of token position i, respectively. These are normalized as:

$$\text{Saliency}_i^{\text{char,LRP}} = \frac{\max(0, R_i^{\text{char}})}{\max_j R_j^{\text{char}}}, \qquad \text{Saliency}_i^{\text{word,LRP}} = \frac{\max(0, R_i^{\text{word}})}{\max_j R_j^{\text{word}}}$$

The two contributions are then combined with clipping, consistent with Eq. (3.6):

$$\text{Saliency}_i^{\text{LRP}} = \texttt{clip}\Big(\text{Saliency}_i^{\text{char,LRP}} + \text{Saliency}_i^{\text{word,LRP}},\, 0,\, 1\Big) \tag{3.11}$$

[1] https://github.com/rachtibat/LRP-eXplains-Transformers

This formulation ensures comparability with other attribution methods (e.g., IG, SeqIG, LIME, SHAP) and enables consistent token-wise visualizations across experiments.

3.7 Context-Aware Multi-Layer Embedding Attribution

To incorporate information from multiple depths of Ithaca's transformer and yield more context-sensitive attributions, we extract token embeddings from several intermediate layers and aggregate their contributions. Concretely, let

$$x_i^{(\ell)} \in \mathbb{R}^D, \quad \ell = 0, 1, \ldots, L$$

be the embedding of token i after transformer layer ℓ (with $\ell = 0$ the input embedding and $\ell = L$ the final embedding). We compute an attribution score at each level:

$$A_i^{(\ell)} = \left\| x_i^{(\ell)} \odot \frac{\partial F(\mathbf{x}^{(\ell)})}{\partial x_i^{(\ell)}} \right\|_2 \quad \text{for } \ell = 0, \ldots, L$$

These per-layer attributions measure how perturbations at different depths affect the final prediction. We then fuse them into a single context-aware saliency score by a weighted sum:

$$\mathrm{A_i} = \sum_{\ell=0}^{L} w_\ell \, A_i^{(\ell)}, \quad \sum_{\ell=0}^{L} w_\ell = 1,$$

where $A_i^{(\ell)}$ denotes the attribution of token i at layer ℓ, and w_ℓ are non-negative layer weights.

In our experiments, we examine embeddings from one transformer layer at a time. This corresponds to setting

$$w_\ell = \begin{cases} 1 & \text{if } \ell = \ell^\star \text{ (chosen layer)} \\ 0 & \text{otherwise,} \end{cases}$$

Finally, we normalize across tokens. This is computed separately for character- and word-level embeddings:

$$\text{Saliency}_i^{\text{char,ML}} = \frac{A_i^{\text{char}}}{\max_j A_j^{\text{char}}}, \qquad \text{Saliency}_i^{\text{word,ML}} = \frac{A_i^{\text{word}}}{\max_j A_j^{\text{word}}},$$

and the two contributions are then combined with clipping, consistent with Eq. (3.6):

$$\text{Saliency}_i^{\text{ML}} = \texttt{clip}\Big(\text{Saliency}_i^{\text{char,ML}} + \text{Saliency}_i^{\text{word,ML}},\, 0,\, 1\Big)\,. \tag{3.12}$$

By aggregating gradients at each intermediate representation, this method captures not only the final model's sensitivity to each token, but also how earlier contextualized embeddings shape the decision.

4 Data

This thesis utilizes a filtered subset of the *Onomastics* dataset, originally introduced by Assael et al. (2022) as part of the *Ithaca* project [Ass+22]. The dataset is a curated subset of the *Packard Humanities Institute* (PHI) Greek Inscriptions corpus (I.PHI)[1], explicitly designed to investigate the relationship between Greek personal names and the chronological attribution of ancient inscriptions. It enables the evaluation of name-based attribution baselines and provides a rigorous benchmark for assessing the predictive utility of personal names in dating historical texts [Ass+22]. The following sections will provide a detailed overview of the *Onomastics* dataset, along with an exploratory data analysis to fully understand its characteristics.

4.1 Onomastics dataset

Each inscription contains at least one personal name that could be matched to an entry in the *Lexicon of Greek Personal Names* (LGPN)[2]. The LGPN provides temporal metadata by associating each name with a probability distribution over 160 ten-year chronological bins spanning the period from 800 BCE to 800 CE [Par19].

Each entry in the Onomastics dataset includes the following information:

- the **PHI identifier** of the inscription, as recorded in the I.PHI corpus
- the **text** of the inscription
- the set of **recognized personal name n-grams**, typically bigrams, identified via LGPN
- the **LGPN-derived date distribution**, computed by aggregating the individual name-level histograms
- the **ground-truth chronological label**, defined as the midpoint of the PHI-provided date interval, discretized into the nearest ten-year bin.

[1] https://inscriptions.packhum.org/
[2] https://www.lgpn.ox.ac.uk/

To create a dedicated evaluation testbed for our explainability methods, we extracted from the Onomastics dataset all inscriptions containing at least one "low-variance" bigram, a proper name bigram whose LGPN-derived date distribution is sharply peaked in a single ten-year bin. This yielded approximately 200 inscriptions that serve as ground-truth anchors. An ideal post-hoc explanation should concentrate most of its attribution mass on the known bigram. In addition to LGPN metadata and name-based date histograms, the Onomastics dataset also includes a list of *targeted n-grams* per inscription. These act as attribution anchors in interpretability experiments, such as saliency map analyses [Ass+22].

The subset extracted from the *Onomastics* dataset thus forms the empirical foundation for the chronological attribution and interpretability experiments presented in Chapter 5. While previous work [Ass+22] established that personal names encode strong chronological signals, our aim is not to re-test this hypothesis. Instead, we leverage this property to create a controlled testbed: a faithful attribution method should allocate high saliency scores to the targeted name bigrams.

4.2 Exploratory Data Analysis

To further understand the structure and properties of the subset extracted from the *Onomastics* dataset, we performed an exploratory data analysis focused on the frequency and distribution of Greek personal name n-grams to capture common morphological patterns (e.g., suffixes, prefixes, or name pairs) that are informative for historical and chronological variation. The results provide valuable insight into the patterns of name usage and their potential value for chronological attribution. Although the dataset refers to *n-grams* in general, we empirically observed that all attested name sequences in our filtered split were exclusively *bigrams*. No higher-order (e.g., trigram or beyond) personal name constructions appeared with sufficient frequency to warrant inclusion.

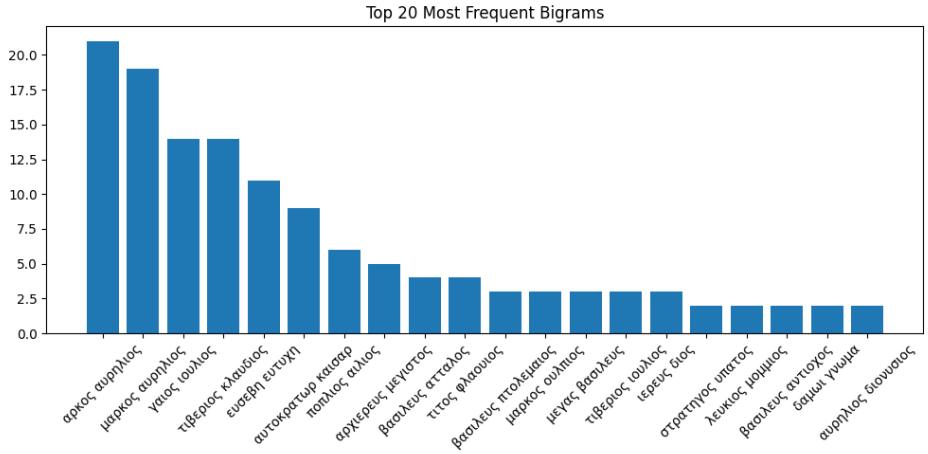


Fig. 4.1: Top 20 most frequent personal name bigrams in the Onomastics subset. The vertical axis indicates the frequency of each bigram.

Figure 4.1 illustrates the 20 most frequently occurring bigrams within the dataset. Notably, names such as ‘αρκος αυρηλιος’, ‘μαρκος αυρηλιος’, and ‘γαιος ιουλιος’ appear with high frequency. These results reflect the dominance of Roman naming conventions, particularly in inscriptions dated to the Imperial period. Such patterns are especially prevalent in regions like Egypt and Asia Minor, and they confirm that certain onomastic patterns can serve as strong indicators of specific historical periods [Sal94].

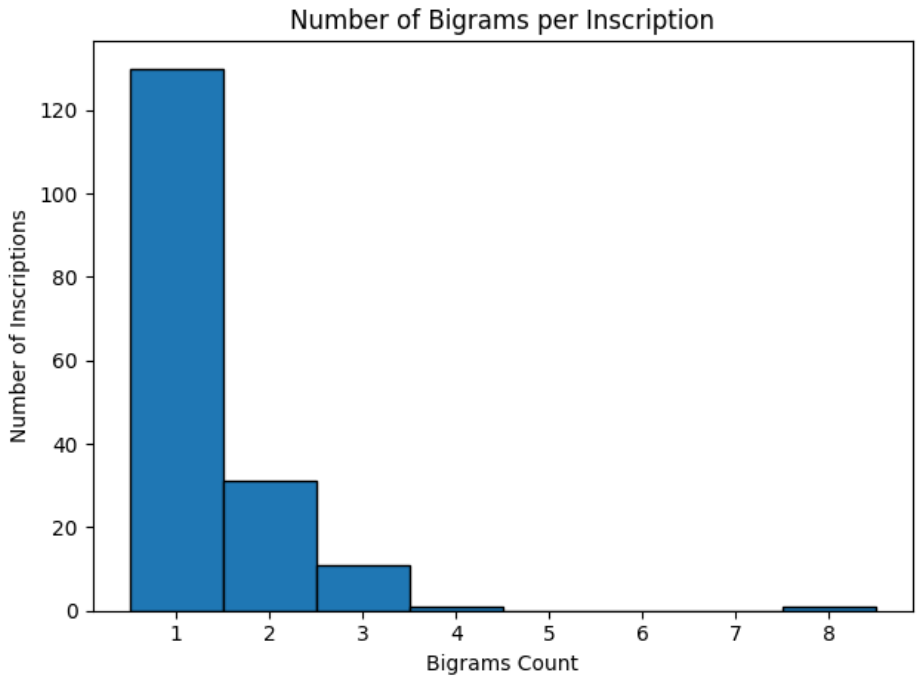


Fig. 4.2: Distribution of the number of the ground-truth bigrams per inscription.

The histogram in Figure 4.2 shows the distribution of the number of the ground-truth bigrams per inscription. The majority of inscriptions contain only one or two LGPN-recognized bigrams. This sparsity highlights the need for models that can make accurate chronological predictions even with minimal input data. Moreover, it motivates the use of interpretable methods such as saliency maps to identify which limited tokens contribute most to the model’s prediction [Ass+22].

Techniques such as saliency maps and attribution scores can help identify which individual bigrams among the sparse input exert the most decisive influence on the predicted temporal label, providing insights into the underlying historical regularities that the model has internalized [Ass+22][STY17] [Sam+21].

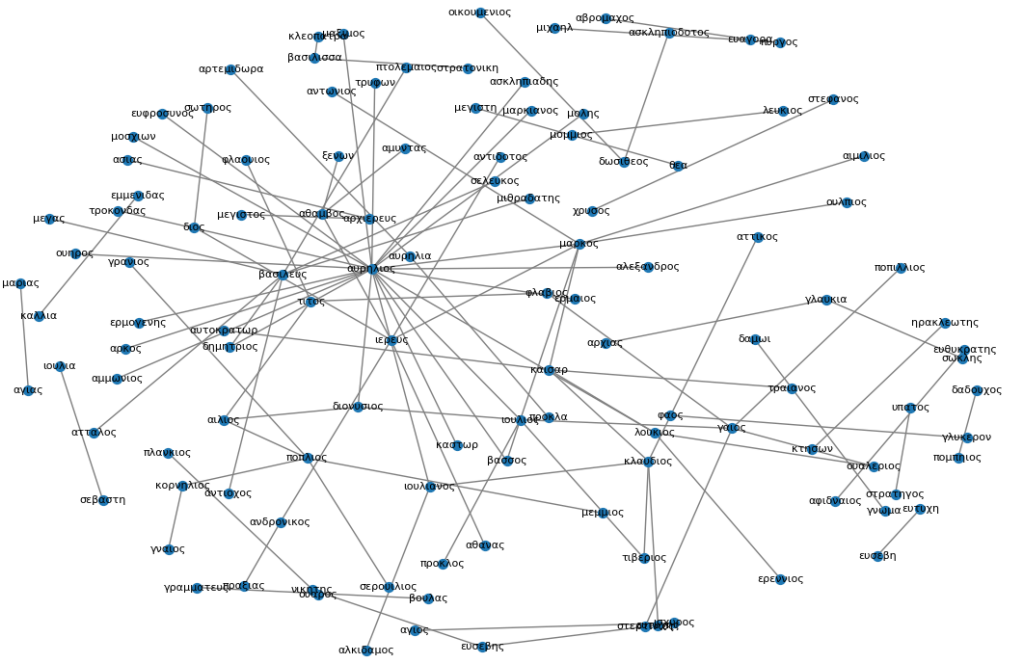


Fig. 4.3: Graph-based visualization of bigram co-occurrence. Nodes represent individual names and edges denote co-occurrence in a bigram.

To examine the co-occurrence structure of names, Figure 4.3 presents a graph-based visualization where nodes correspond to individual words used in person name bigrams and edges denote co-occurrence in at least one person-name bigram. While the network's complexity and density obscure individual node labels in this overview, the overall structure reveals important patterns in name relationships. The graph exhibits a hub-and-spoke topology, where certain names act as central connectors with high degree centrality, indicating their frequent pairing with diverse other names. Most prominently, highly connected nodes such as 'αυρηλιος' (Aurelius) and 'βασιλευς' (basileus, king) emerge as central hubs, reflecting the prevalence of Roman naming conventions and imperial titles in the dataset. The clustering patterns visible in the network suggest that certain names exhibit temporal or regional regularities in their pairings. Roman naming elements—including praenomina (personal names) and nomina (family names)—tend to cluster together (e.g., 'γαιος ιουλιος' (Gaius Julius), 'μαρκος αυρηλιος' (Markos Aurelius)), while Greek names form distinct sub-networks (e.g., 'δημήτριος σωκράτης' (Demetrios Sokrates), 'αντίοχος νικόλαος' (Antiochos Nikolaos)). This network structure reveals the underlying onomastic patterns in ancient Greek inscriptions and could be exploited in future approaches leveraging graph-based representation learning for attribution tasks [HYL17].

5 Experiments and Results

The central aim of this chapter is to evaluate a series of explainability methods applied to the model predictions for the task of chronological attribution of ancient Greek inscriptions using the Onomastics subset. Specifically, we aim to answer the following research questions: (1) Which attribution methods most accurately and faithfully highlight the historically relevant tokens, i.e., the proper-name bigrams? (2) How do design decisions in explanation—such as how attribution scores are aggregated, at what representation level they are computed (characters or words), and how contextual information is incorporated—affect the reliability of these explanations? These questions are motivated by recent developments in neural interpretability, which suggest that both the choice of attribution method and the internal representations of the model critically shape the quality of the explanation [DeY+20] [Val+23].

5.1 Assumptions and Scope

Throughout this chapter we adopt two assumptions for the Onomastics subset: (i) the annotated proper–name bigrams are the most informative features for chronological attribution; and (ii) for *correctly classified* inscriptions, Ithaca indeed relies on these bigrams to reach its decisions. Under these assumptions, an attribution method that faithfully reflects the model's internal reasoning should highlight the person–name tokens. We therefore evaluate methods by the extent to which their saliency maps focus on the annotated name bigrams. We note, however, that this criterion can also be read as measuring *plausibility* (alignment with expert expectations) rather than strict *faithfulness* (matching the model's true decision process), especially for *incorrectly classified* cases, where the model may rely on other signals [JG20]. To target faithfulness (rather than plausibility), all primary metrics are computed on the subset of inscriptions that Ithaca classifies correctly, where the assumption that the model uses proper–name bigrams is most reasonable.

5.2 Evaluation Metrics

To quantify how well each attribution method localizes the target bigram, we evaluate saliency maps using both retrieval-style and classification-style metrics.

5.2.1 Retrieval-Style Metrics

We treat the two tokens of the targeted bigram as relevant items and rank all tokens by descending attribution score. We compute the following retrieval metrics:

$$\text{MRR}_{\text{avg}},\ \text{MRR}_{\text{max}},\ \text{MAP}_{\text{avg}},\ \text{MAP}_{\text{max}},\ \text{nDCG@2}_{\text{avg}},\ \text{nDCG@2}_{\text{max}}$$

Here, "avg" denotes the average score over the two bigram tokens, while "max" reports the highest score of the two. This dual view captures both overall alignment and best case localization.

Mean Reciprocal Rank (MRR)
MRR quantifies how early a relevant token (i.e., a token belonging to the target bigram) appears in the ranked saliency list for each inscription. In this context, each query corresponds to one evaluated inscription. MRR is defined as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{5.1}$$

where $|Q|$ is the number of evaluated inscriptions (queries), and rank_i is the position of the first relevant token (from the target bigram) in the saliency-based ranking for inscription i. A higher MRR indicates that relevant tokens are ranked closer to the top on average [Voo+99].

Mean Average Precision (MAP)
MAP evaluates the ability of the saliency ranking to place the relevant tokens of each inscription (i.e., the annotated bigram) near the top of the ranking. For inscription i, the *Average Precision* (AP) is defined as the mean of the Precision values at the ranks where relevant tokens occur:

$$\text{AP}_i = \frac{1}{m_i} \sum_{k=1}^{n_i} P_i(k) \cdot \text{rel}_i(k), \tag{5.2}$$

where m_i is the number of relevant tokens in inscription i ($m_i = 2$ for bigrams), n_i is the total number of tokens, and $\text{rel}_i(k) = 1$ if the token at rank k belongs to the annotated bigram and 0 otherwise. The term $P_i(k)$ denotes the Precision at rank k, formally defined as

$$P_i(k) = \frac{1}{k} \sum_{j=1}^{k} \text{rel}_i(j). \tag{5.3}$$

Mean Average Precision (MAP) is then obtained by averaging AP over all inscriptions:

$$\text{MAP} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \text{AP}_i. \tag{5.4}$$

MAP therefore reflects both (i) how highly the relevant tokens are ranked and (ii) whether all relevant tokens are successfully retrieved [Voo+99].

Normalized Discounted Cumulative Gain at 2 (nDCG@2)
Normalized Discounted Cumulative Gain (nDCG) is a ranking-based metric that evaluates how highly relevant tokens appear within the saliency ranking of each inscription. For inscription i, the Discounted Cumulative Gain at cutoff k is defined as

$$\mathrm{DCG}_i@k = \sum_{j=1}^{k} \frac{2^{\mathrm{rel}_i(j)} - 1}{\log_2(j+1)}, \tag{5.5}$$

where $\mathrm{rel}_i(j) \in \{0, 1\}$ indicates whether the token at rank j belongs to the annotated ground-truth bigram. The Ideal DCG (IDCG) is computed analogously, but with the relevant tokens placed in the highest possible ranks. The normalized score is then

$$\mathrm{nDCG}_i@k = \frac{\mathrm{DCG}_i@k}{\mathrm{IDCG}_i@k}, \tag{5.6}$$

and the final result is the mean nDCG@k across all inscriptions.

In our setting, most inscriptions contain a single annotated bigram, i.e., two relevant tokens. Using $k = 2$ is therefore the most informative choice in the typical case: nDCG@2 directly measures whether both tokens of the bigram are concentrated among the top two ranks. When both tokens appear in the first two positions, nDCG@2 reaches its maximum value of 1.0; if only one or none appear, the score is lower, reflecting incomplete retrieval of the bigram. For inscriptions with multiple annotated bigrams, nDCG@2 still provides a consistent top-rank evaluation, albeit underestimating the attainable gain. This makes nDCG@2 a natural complement to MAP, as it emphasizes top-rank accuracy while still accounting for relevant tokens [JK02].

Retrieval-style metrics operate at the token level: all tokens in an inscription are ranked, and the two tokens of the annotated bigram constitute the relevant set. By contrast, classification-style metrics (next section) operate at the bigram level: token scores are pooled into span scores before thresholding.

5.2.2 Classification-Style Metrics

To complement retrieval-based evaluation at the token level, we also adopt a binary classification framework that evaluates attribution at the bigram level. Here, token-level attribution scores are first aggregated into bigram-level values (using the pooling schemes introduced in Section 5.3.1). We then threshold the bigram scores at the 90th percentile: specifically, the top 10% of bigrams with the highest attribution values are labeled as "selected" (*positives*), while the rest are considered "not selected" (*negatives*). These predic-

tions are compared to the annotated ground-truth bigram span to compute the following standard metrics:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5.7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5.8}$$

$$\text{F1} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.9}$$

where TP (true positives) is the number of ground-truth bigrams correctly selected, FP (false positives) is the number of non–ground-truth bigrams incorrectly selected, and FN (false negatives) is the number of ground-truth bigrams missed by the selection.

This evaluation follows standard rationale selection protocols [DeY+20]. The 10% threshold is a convention commonly used to compensate for differences in saliency value scales across methods. [DeY+20].

To complement threshold-based metrics and assess the discrimination quality of saliency rankings regardless of any specific threshold, we also report the area under the Precision–Recall curve:

$$\text{AUC}_{\text{PR}} = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}) \tag{5.10}$$

A higher AUC_{PR} indicates that the method more successfully separates relevant (ground-truth) from irrelevant bigrams across all thresholds, providing a threshold-independent estimate of attribution fidelity [DeY+20].

Note: In our experiments, "selected" bigrams refer to contiguous token pairs, with ground truth defined by the annotated target bigram for each inscription. All metrics are macro-averaged over the full test set (Onomastics subset).

5.3 Evaluation

5.3.1 Experiment 1: Evaluating Aggregation Bias and Saliency Granularity

Before comparing explainability methods on the full testbed, we conducted a preliminary study to determine optimal aggregation strategies for token-level saliency scores. This experiment addresses two key design choices: **(a)** how to pool saliency scores across bigram

spans, and **(b)** whether to use character-level, word-level, or combined embeddings. Although attribution methods operate at the token level, our evaluation benchmark defines ground truth at the bigram level. To ensure consistency with the annotations, we apply evaluation metrics to bigram units rather than individual tokens. This requires pooling token-level attribution scores into a single bigram-level score (using sum, max, or average). Thus, pooling is not meant to alter the attribution methods themselves, but to align their token-level outputs with the span-level ground truth of our benchmark.

Experimental Setup

For each inscription, after computing the token-level attribution scores using the baseline method (Gradient $\odot$ Input) described in Chapter 3, we computed a single bigram saliency score by pooling the token scores across the span of each candidate bigram (e.g., a proper name). Specifically, for a bigram consisting of token indices $S = \{i_1, i_2\}$, we tested the following pooling schemes:

1. **Sum:** $\text{Saliency}_{\text{bigram}} = \sum_{i \in S} \text{Saliency}_i$
2. **Max:** $\text{Saliency}_{\text{bigram}} = \max_{i \in S} \text{Saliency}_i$
3. **Average:** $\text{Saliency}_{\text{bigram}} = \frac{1}{|S|} \sum_{i \in S} \text{Saliency}_i$

We also compared three levels of saliency computation:

- **Character embeddings only:** per character saliency map
- **Word embeddings only:** per word saliency map
- **Character + word embeddings:** saliency scores are first normalized separately for character- and word-level embeddings, then summed and clipped to the $[0, 1]$ range to produce a combined, balanced saliency score.

Evaluation Metrics

We measured retrieval-style metrics (MRR, MAP, nDCG@2) on our *low-variance bigram* test set, treating the annotated proper name bigrams as the relevant items and ranking all bigrams by their pooled saliency.

Results: Aggregation and Granularity

Figure 5.1 shows that: **Sum** pooling outperforms both **max** pooling and **average** pooling in most cases across MRR, MAP, and nDCG@2 and **word** level is better than **char** or

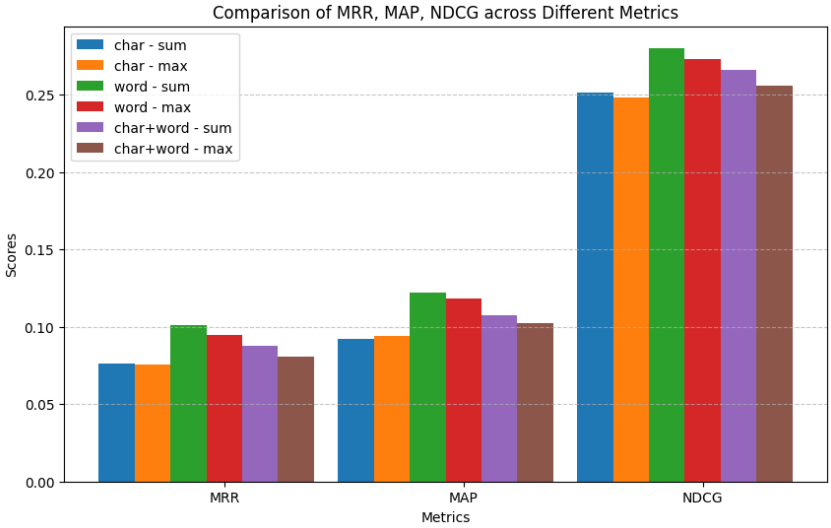


Fig. 5.1: Comparison of MRR, MAP and nDCG@2 for all six combinations of embedding granularity—character-level, word-level, and combined character plus word embeddings—and pooling strategies, specifically sum pooling and max pooling.

word plus char level. The average pooling results are omitted from the plot to avoid visual clutter, as they were slightly below the max pooling scores.

Robustness Check: Removing Non-Word Tokens

After establishing the relative effectiveness of different aggregation and granularity strategies (Fig. 5.1), we further examined the robustness of our chosen approach. Specifically, we wished to ensure that the observed differences were not artifacts introduced by irrelevant tokens, such as spaces, UNK tokens, or punctuation marks. To this end, we repeated the aggregation analysis for the three sum-based granularities (character, word, char+word), but *after zeroing out all non-word tokens* in the saliency maps.

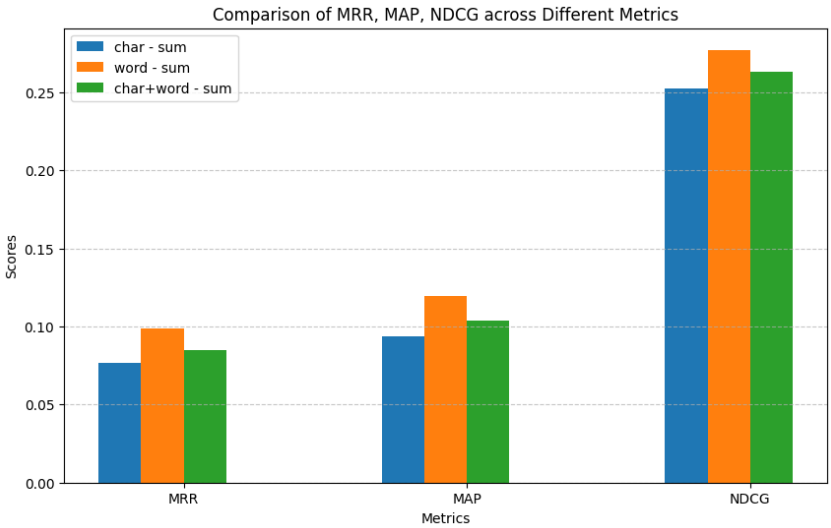


Fig. 5.2: Ablation study repeated after excluding non-word tokens (e.g., punctuation or special markers) from the saliency computation. The sum strategy continues to perform best.

As shown in Fig. 5.2, excluding non-word tokens does *not* materially affect the ranking of aggregation schemes. So sum aggregation continues to outperform max and avg (results for avg not shown). We adopt char+word saliency for all further experiments, because this is consistent with the original Ithaca implementation, which leverages both character- and word-level information, even though word only is better.

Bias Analysis: Bigram Length Effects

However, we note a potential source of bias with sum-based aggregation: the proper names targeted as bigrams in our evaluation are typically longer in characters than other words in the corpus. This length difference could artificially inflate bigram saliency scores under sum pooling. To allay concerns about bias from name bigrams being longer tokens, we compared the length distributions of average char-length of the single words of the target bigrams and the char-length of the average word:

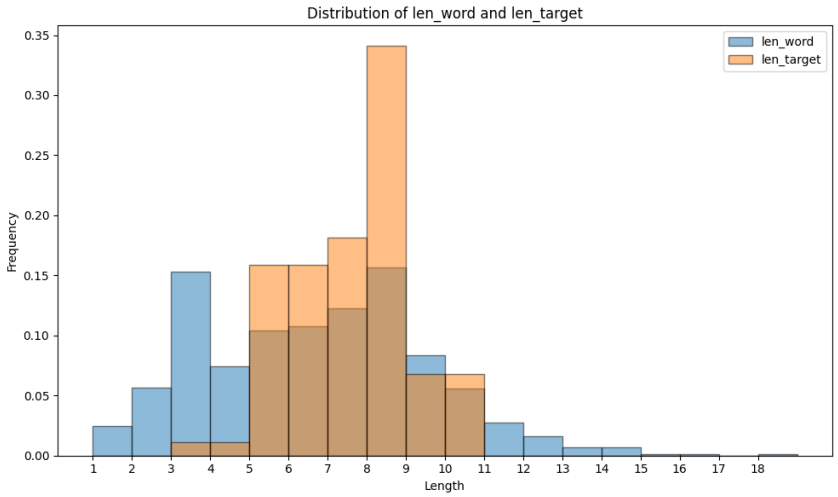


Fig. 5.3: Histogram of token character lengths: blue = all words, orange = words of targeted bigrams. Targeted bigrams are longer, which can bias a simple sum.

As illustrated in Fig. 5.3, average char-length of the single words of the target bigrams is larger than the char-length of the average word in general. This indeed introduces a potential bias for sum aggregation, as longer spans are likely to accumulate higher total saliency scores simply by having more characters. To mitigate this bias and ensure fair evaluation of saliency localization, we therefore decided to study max and average pooling strategies. These approaches reduce the direct dependence of the saliency score on bigram length, allowing for a more reliable comparison of attribution methods across variable-length entities.

Final Decision: Max Pooling

Ultimately, we adopted max pooling, which assigns to each bigram the highest saliency value among its constituent tokens.

This choice is motivated by the following considerations:

- **Length invariance:** Max pooling ensures that longer bigrams are not unfairly favored, as only the most salient token determines the bigram's score [DeY+20].

- **Sensitivity to key tokens:** In chronological attribution of ancient texts, a single highly distinctive name can be sufficient for confident dating [Ass+22]. Max pooling is especially sensitive to such sharply focused attributions.

- **Alignment with expert intuition:** Qualitative analysis suggests that expert annotators often rely on the most distinctive token within a name. Max pooling more faithfully reflects this "peak evidence" approach [DeY+20] [JG20].

- **Noise considerations:** While max pooling can be sensitive to spurious spikes, we observed that saliency maps for proper names tend to be robust in practice.

While sum-pooling of char+word attributions initially yielded the best retrieval scores, we observed a significant length bias due to the longer average length of target bigrams. To ensure fair comparison across variable-length names, we therefore adopt max pooling for all further experiments. This decision allows each bigram's most salient token to determine its overall score, preventing spurious inflation of importance for longer names. This approach is consistent with best practice in rationale selection evaluation [DeY+20].

5.3.2 Experiment 2: Method Comparison and Ablations

Having established in Experiment 1 (see Sec. 5.3.1) the most robust pooling and aggregation strategies, we now systematically compare state-of-the-art attribution methods for their ability to correctly localize relevant bigrams. This experiment aims to identify which techniques provide the most faithful token-level explanations, serving as a foundation for further context-aware analyses in Experiment 3 (see Sec. 5.3.3).

Explainability Techniques Evaluated
The following explainability techniques are benchmarked:

- **Gradient:** The raw gradient of the model output with respect to input embeddings (see Sec. 3.1).

- **Gradient $\odot$ Input:** The elementwise product of input embeddings and output gradients (see Sec. 3.1).

- **Integrated Gradients (IG):** Four variants are considered (see Sec. 3.2):

  - Multi-step IG with zero baseline

  - Multi-step IG with input centroid baseline
  - Single-step IG with zero baseline
  - Single-step IG with input centroid baseline

- **Sequential Integrated Gradients (SeqIG):** IG applied sequentially, interpolating one token at a time, holding others fixed (see Sec. 3.3).
- **Local Interpretable Model-agnostic Explanations (LIME):** A surrogate-based method using local linear approximations of the model around each input (see Sec. 3.4).
- **SHapley Additive exPlanations (SHAP):** Assigns token-level attributions based on Shapley values from cooperative game theory, using Kernel SHAP for tractability (see Sec. 3.5).
- **Layer-wise Relevance Propagation (LRP):** Redistributes model output relevance back to input tokens (see Sec. 3.6).
- **Attention-Aware LRP (AttnLRP):** A transformer-specific extension of LRP that propagates relevance through multi-head self-attention using normalized attention coefficients (see Sec. 3.6).

Experimental Protocol

For each inscription, we first compute per-token saliency scores using the selected attribution method. For classification-style metrics (Sec. 5.2), we then obtain bigram-level scores by max pooling the token-level values:

$$\text{Saliency}_{\text{bigram}} = \max_{i \in \text{bigram}} \text{Saliency}_i$$

This choice avoids length bias and reflects the intuition that a single highly salient token can suffice for attribution (see Experiment 1 – Sec. 5.3.1). Retrieval-style metrics, by contrast, are computed directly on the token-level scores, treating the two tokens of the annotated bigram as the relevant set. In all cases, token-level saliency is defined as the sum of character- and word-level contributions (*char+word*).

Results

Table 5.1 reports the performance of each explainability method. Integrated Gradients with input centroid baseline, randomized multi-step integration and the Sequential IG yield the highest MRR and MAP, as well as nDCG@2, indicating superior Precision and

Tab. 5.1: Comparative evaluation of token-level explainability methods (char+word). For each retrieval metric (MRR, MAP, nDCG@2), we report both *average* performance (scores averaged across the two ground-truth tokens of each bigram) and *best-case* performance (score of the higher-ranked token). Average scores capture how well a method highlights *both* tokens of the name, while best-case scores reflect whether *at least one* token is strongly emphasized. We report both *single* variants (a simple one-step interpolation) and *multi* variants (path integration with $m = 50$ steps), each with either a *zero* or *centroid* baseline. SeqIG achieves the highest values across both perspectives, outperforming IG, LRP, AttnLRP, LIME, and SHAP.

Method	AVG MRR	AVG MAP	AVG nDCG	MAX MRR	MAX MAP	MAX nDCG
Gradient	0.0503	0.0705	0.2001	0.0707	0.1109	0.2104
Grad x Input	0.0813	0.1004	0.2586	0.0992	0.1324	0.2710
IG (zero, multi)	0.1660	0.1943	0.3371	0.2069	0.2370	0.3681
IG (zero, single)	0.0815	0.1003	0.2588	0.0994	0.1326	0.2714
IG (centroid, multi)	0.1908	0.2212	0.3504	0.2203	0.2802	0.3903
IG (centroid, single)	0.1404	0.1702	0.3205	0.1808	0.2001	0.3306
SeqIG (zero, multi)	**0.2290**	**0.2612**	**0.3821**	**0.2669**	**0.3257**	**0.4111**
LRP	0.0902	0.1104	0.2403	0.1005	0.1508	0.2502
AttnLRP	0.1328	0.1641	0.3107	0.1712	0.2157	0.3323
LIME	0.1209	0.1504	0.2908	0.1607	0.1903	0.3106
SHAP	0.1907	0.2208	0.3601	0.2306	0.2808	0.3901

localization of salient spans. Gradient $\odot$ Input remains a competitive and computationally efficient baseline.

Tab. 5.2: Classification-based performance across explainability methods. Each method's bigram-level Precision, Recall, F1, and AUC (area under the Precision–Recall curve) are reported, using a 90th-percentile saliency threshold to select positive bigrams. Higher is better for all metrics. We report both *single* variants (a simple one-step interpolation) and *multi* variants (path integration with $m = 50$ steps), each with either a *zero* or *centroid* baseline.

Method	**Precision**	**Recall**	**F1**	**AUC**
Gradient	0.08	0.17	0.11	0.47
Grad $\times$ Input	0.12	0.22	0.16	0.51
IG (zero, multi)	0.20	0.40	0.27	0.60
IG (zero, single)	0.13	0.21	0.17	0.52
IG (centroid, multi)	0.25	0.50	0.33	0.65
IG (centroid, single)	0.18	0.35	0.24	0.58
SeqIG (zero, multi)	**0.30**	**0.60**	**0.41**	**0.70**
LRP	0.13	0.25	0.17	0.52
AttnLRP	0.16	0.31	0.22	0.55
LIME	0.15	0.30	0.20	0.54
SHAP	0.24	0.48	0.32	0.64

As shown in Table 5.2, the trends are consistent with the retrieval metrics. SeqIG again achieves the best performance with the highest Precision and Recall (about 30% and 60% at the 90th-percentile threshold, respectively, yielding F1$\approx$0.40). This indicates that SeqIG can retrieve a substantial portion of the true spans while keeping false positives relatively low. IG (centroid, multi) and SHAP form the next tier of performance (F1$\approx$0.32–0.33), while the basic gradient-based methods trail behind.

Notably, all methods exhibit moderately low Precision at the chosen threshold despite a decent Recall. For instance, even the best method (SeqIG) only achieves ∼30% Precision, meaning many non-target bigrams are falsely selected. This underscores the difficulty of achieving high Precision and Recall simultaneously for token-level attribution [DeY+20] [JG20]. Nonetheless, SeqIG's superior F1 and AUC highlight its advantage in accurately pinpointing the relevant bigram span compared to other techniques.

5.3.3 Experiment 3: Layer-wise and Multi-layer Contextual Attribution

Building on the findings of Experiment 2 (see Sec. 5.3.2), that SeqIG performs best when applied to the first layer, we next investigate whether examining intermediate or aggregated upper-layer representations can further improve the faithfulness of token-level attributions. To our knowledge, the effect of applying IG or SeqIG to upper-layer hidden states has not been systematically studied before, making this analysis a novel contribution of the present thesis.

Recent research on transformer interpretability has shown that semantic and contextual information peaks in the upper-middle and final layers of the network, often making these layers the most informative for saliency attribution [Ass+22] [JSS19] [RKR21] [TDP19].

Experimental Protocol
In this experiment we use the Context-Aware Multi-Layer Embedding Attribution methods introduced in Chapter 3 (see Sec. 3.7). We benchmark two context-aware attribution strategies along with our best method from Experiment 2 (see Sec. 5.3.2) using our testbed:

1. **Penultimate Layer Attribution:** We apply SeqIG to the **7th (penultimate) transformer layer** of the Ithaca model, following evidence that this layer balances semantic richness and task focus [Ass+22].

2. **Multi-layer Averaged Attribution:** We compute SeqIG at each of the final three layers (6, 7, and 8) and average the resulting token attributions:

$$\text{Saliency}_i^{\text{avg}} = \frac{1}{3}\sum_{\ell=6}^{8} \text{Saliency}_i^{(\ell)}$$

   This approach is motivated by findings that aggregating attributions from multiple late layers can enhance stability and capture the peak of contextual integration [HLL25].

For both strategies, we follow the same pooling and evaluation protocol as in Experiment 2 (see Sec. 5.3.2). Retrieval metrics (MRR, MAP, nDCG@2) and classification metrics (Precision, Recall, F1, AUC) are then calculated over the testbed.

Results

Tab. 5.3: Comparison of token-level attribution performance for SeqIG of Experiment 2 (see Sec. 5.3.2) vs. penultimate layer vs. mean of last three layers (char+word). Best results bolded.

Layer Attribution	AVG MRR	AVG MAP	AVG nDCG	MAX MRR	MAX MAP	MAX nDCG
SeqIG (Layer 1 - Exp 2)	0.2290	0.2612	0.3821	0.2669	0.3257	0.4111
SeqIG (Layer 7)	0.2345	0.2643	0.3861	0.2702	0.3301	0.4185
SeqIG (Mean 6–8)	**0.2421**	**0.2730**	**0.3933**	**0.2796**	**0.3398**	**0.4260**

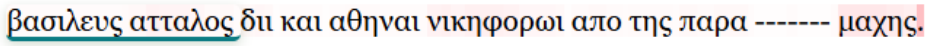
Table 5.3 reports the performance of SeqIG when applied at the penultimate layer versus when averaging the final three layers.

Tab. 5.4: Classification-based attribution performance for SeqIG of Experiment 2 (see Sec. 5.3.2), layer-wise and multi-layer SeqIG methods (char+word). Each method's Precision, Recall, F1, and AUC are reported, using a 90th-percentile saliency threshold. Best results are bolded.

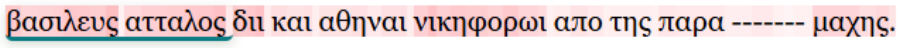
Layer Attribution	Precision	Recall	F1	AUC
SeqIG (Layer 1 - Exp 2)	0.30	0.60	0.40	0.70
SeqIG (Layer 7)	0.32	0.62	0.43	0.72
SeqIG (Mean 6–8)	**0.34**	**0.65**	**0.45**	**0.74**

Our results show that both strategies outperform input-layer attribution (Layer 1) across all metrics in our testbed. Averaging across Layers 6–8 yields the best overall scores, suggesting that aggregating late layer signals can enhance saliency localization. These findings are in line with prior evidence that upper-layer representations encode richer contextual signals [HLL25] [Ass+22], and they highlight the practical value of multi-layer, context-aware attribution in epigraphic NLP. We emphasize that we did not evaluate every individual upper layer; thus, we refrain from claiming that all upper layers (e.g., Layers 6 or 8) necessarily outperform Layer 1. We also leave a systematic comparison of early- and mid-layer averages (e.g., Mean 1–3, Mean 3–5) to future work.

To complement our quantitative results, we present qualitative visualizations of token-level saliency maps for representative inscriptions. These heatmaps illustrate how different attribution methods highlight the tokens deemed most relevant for chronological attribution. In particular, we compare the best-performing context-aware method (Sequential Integrated Gradients, mean of layers 6–8) with the baseline Gradient $\odot$ Input. As shown above, sharper and more focused highlighting of the ground-truth bigram provides a more faithful and interpretable explanation.

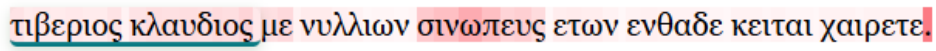


(a) Token-level saliency map for the inscription using the baseline method (Gradient ⊙ Input). The relevant tokens are less distinctly highlighted and other tokens receive spurious attribution.

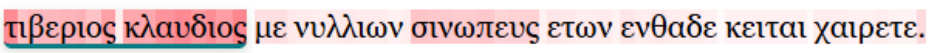


(b) Token-level saliency map for the inscription using the best-performing method (SeqIG, mean of layers 6–8). The correct bigram 'βασιλευς ατταλος' is strongly highlighted, indicating precise localization of relevant tokens.

Fig. 5.4: Illustration of token-level attribution for a representative inscription. The best method (SeqIG, mean of layers 6–8) yields sharper and more focused saliency on the ground-truth bigram compared to the baseline, aligning better with historical expectations.



(a) Token-level saliency map for the inscription using the baseline method (Gradient ⊙ Input). The relevant tokens are less distinctly highlighted and other tokens receive spurious attribution.



(b) Token-level saliency map for the inscription using the best-performing method (SeqIG, mean of layers 6–8). The correct bigram 'τιβεριος κλαυδιος' is strongly highlighted, indicating precise localization of relevant tokens.

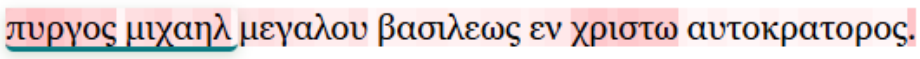
Fig. 5.5: Illustration of token-level attribution for a representative inscription. The best method (SeqIG, mean of layers 6–8) yields sharper and more focused saliency on the ground-truth bigram compared to the baseline, aligning better with historical expectations.

In addition to the representative cases above, we provide 2 further examples to illustrate the strengths and limitations of the attribution methods. While Sequential Integrated Gradients (mean of layers 6–8) typically produces sharper and more localized highlighting of the ground-truth bigram compared to the baseline Gradient ⊙ Input, this pattern is not universal.

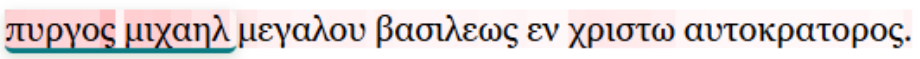
5.4 Discussion

The results from our experiments yield several important insights into the effectiveness of different attribution strategies for saliency localization in the context of ancient Greek onomastic data.

First, our preliminary study on aggregation and granularity (Experiment 1, Sec. 5.3.1) demonstrated that sum-based pooling consistently yields the highest retrieval metrics across most settings, especially when combining character- and word-level saliency. However, we identified a systematic bias in sum-based pooling: since the proper names used as bigram targets are typically longer than average words, their saliency scores can be artificially inflated under sum aggregation. To mitigate this, we adopted max pooling in

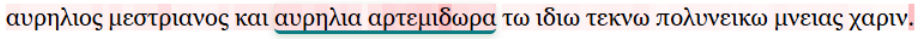


(a) Token-level saliency map for the inscription using the baseline Gradient ⊙ Input. The correct bigram 'πυργος μιχαηλ' is more sharply highlighted compared to Sequential Integrated Gradients (mean of layers 6–8), indicating that in this example the simpler baseline provides better localization of the relevant tokens.

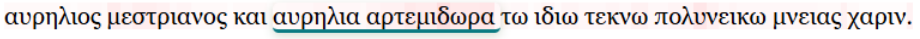


(b) Token-level saliency map for the inscription using the best-performing method (SeqIG, mean of layers 6–8). The relevant tokens are less distinctly highlighted

Fig. 5.6: Illustration of token-level attribution for a representative inscription. The baseline yields sharper and more focused saliency on the ground-truth bigram compared to the best method (SeqIG, mean of layers 6–8).



(a) Token-level saliency map for the inscription using the baseline Gradient ⊙ Input. The correct bigram 'αυρηλια αρτεμιδωρα' is more sharply highlighted compared to Sequential Integrated Gradients (mean of layers 6–8), indicating that in this example the simpler baseline provides better localization of the relevant tokens.



(b) Token-level saliency map for the inscription using the best-performing method (SeqIG, mean of layers 6–8). The relevant tokens are less distinctly highlighted

Fig. 5.7: Illustration of token-level attribution for a representative inscription. The baseline yields sharper and more focused saliency on the ground-truth bigram compared to the best method (SeqIG, mean of layers 6–8).

all subsequent experiments, ensuring length-invariance and focusing evaluation on the most informative token within each bigram. This approach is further supported by qualitative evidence: expert epigraphers often identify a single key token as the decisive clue for dating an inscription.

In our comparative evaluation of attribution methods (Experiment 2, Sec. 5.3.2), we found that gradient-based methods with integration, particularly Sequential Integrated Gradients (SeqIG), significantly outperform simpler approaches such as raw gradients and Gradient ⊙ Input. These findings are in line with recent Transformer interpretability literature [Deh25] [Eng23], which consistently show that IG-based methods, especially with sequential or multi-step integration, achieve the highest fidelity in ranking truly important tokens. In our benchmark, SeqIG delivered the top scores on all retrieval metrics (MRR, MAP, nDCG) and all classification metrics (Precision, Recall, F1, AUC), while Multi-step IG with centroid baseline, AttnLRP and SHAP also performed strongly. Model-agnostic methods such as LIME and SHAP, while robust and easy to apply, were slightly less precise than advanced IG variants, especially for complex or long inscriptions. Layer-wise Rele-

vance Propagation (LRP) provided moderate improvements over simple gradients but did not match the performance of the best IG-based approaches, echoing theoretical observations that LRP and Gradient $\odot$ Input can be equivalent in certain architectures [WO21].

Another essential consideration is computational efficiency. While multi-step and sequential variants of Integrated Gradients (IG), such as SeqIG, achieved the highest overall attribution performance, we observe that single-step IG methods, those using a centroid or zero baseline, still yield competitive results across both retrieval-style and classification metrics. For instance, IG (centroid, single) achieved an F1 score of 0.24 and an AUC of 0.58, along with MAP and MRR scores that surpass several more complex methods. Given that these single-step variants require only a single backward pass, they offer a substantial reduction in computational cost relative to multi-step or layer-wise techniques. This makes them a cost-effective and scalable choice in settings where explanation latency or model interrogation budget is constrained, without severely sacrificing attribution quality.

Building on these findings, Experiment 3 (see Sec. 5.3.3) explored whether the choice of layer for attribution—either focusing on the penultimate (7th) transformer layer or averaging the final three layers—could further improve performance. The motivation here, as supported by [Val+23] [Ass+22] [HLL25], is that contextual and semantic richness often peaks in the upper-middle or final layers of Transformer models. Our results confirm this: both strategies outperform single-layer (last layer) attribution, with the mean of layers 6–8 achieving the highest overall scores. This suggests that integrating information from several upper layers captures a more robust and context-aware signal, yielding more faithful token-level attributions. Notably, the difference in performance between penultimate-layer and multi-layer averages, while present, is moderate, indicating that either approach offers a clear improvement over relying solely on the first layer.

These experiments illustrate the value of using max-pooled, context-aware attributions based on advanced IG techniques and highlight the importance of appropriately leveraging the representational hierarchy of Transformer models. Our best-performing approach (SeqIG with multi-layer aggregation) provides interpretable saliency maps that closely align with ground-truth rationales, supporting robust, historian-friendly model explanations in the challenging domain of ancient text attribution.

Overall, our results demonstrate that state-of-the-art explainability methods, when used with careful attention to model internals and task-specific constraints, can provide meaningful, contextually grounded explanations for transformer-based models applied to ancient text.

6 Conclusions and Future Work

6.1 Conclusions

This thesis systematically investigated explainability methods for transformer-based models applied to the attribution of ancient Greek inscriptions. The study addressed the critical challenge of providing transparent and faithful token-level explanations in digital epigraphy, with a focus on Ithaca model and the Onomastics dataset.

Through a series of carefully designed experiments, we benchmarked a wide spectrum of post-hoc attribution techniques including gradients, Gradient $\odot$ Input, various Integrated Gradients (IG) variants, Sequential Integrated Gradients (SeqIG), model-agnostic methods (LIME, SHAP), Layer-wise Relevance Propagation (LRP), and context-aware multi-layer attribution strategies. Our main findings are summarized as follows:

- **Pooling and granularity choices are critical.** Initial preliminary studies revealed that sum-based pooling can introduce length bias in proper-name bigrams: max pooling provides length invariance and sharper interpretability, while combining character- and word-level saliency yields the most informative token attributions.

- **Advanced IG-based methods deliver the highest fidelity.** Sequential Integrated Gradients (SeqIG), especially when applied to later layers, consistently outperformed baseline methods in retrieval-style metrics (MRR, MAP, nDCG) and classification metrics (Precision, Recall, F1, AUC), confirming results from recent literature. Multi-step IG with centroid baseline and SHAP also proved robust across diverse examples.

- **Layer-wise and multi-layer attributions improve context sensitivity.** Aggregating attributions from the penultimate and last transformer layers further enhanced performance, with the mean of layers 6–8 yielding the best results overall. This aligns with the state-of-the-art understanding that semantic information peaks in upper-middle transformer layers.

Collectively, these findings establish a rigorous pipeline for interpretable neural modeling in the digital humanities and validate that modern explainability techniques can offer robust, historian-friendly insight into the workings of large transformer models like Ithaca.

6.2 Future Work

While this thesis establishes a rigorous baseline for explainability in transformer-based epigraphic NLP, several avenues remain open for advancing the state of the art:

- **Exploration of more sophisticated attribution techniques.** Future research should investigate the use of advanced explainability methods such as DeepSHAP [CLL19], DIG (Discretized Integrated Gradients) [SR21], and contrastive attribution techniques (e.g., Contrast-CAT) [Jac+21], which have shown promising results in recent Transformer benchmarks. Incorporating path-aware or layer-selective attributions could further enhance fidelity.

- **Task-adaptive and dynamically weighted attribution.** Rather than relying on static or uniform aggregation across layers, developing methods that learn optimal layer weights or dynamically adapt attribution strategies based on the type of inscription or specific downstream task could yield even sharper, context-sensitive explanations.

Taken together, this thesis demonstrates that state-of-the-art attribution methods—notably those leveraging context-aware, multi-layer representations—can make transformer models like Ithaca both transparent and trustworthy for historical text analysis. Further research at the intersection of explainable AI and digital humanities promises not only better models, but deeper understanding of the ancient world.

Bibliography

[Ach+24] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, et al. "Attnlrp: attention-aware layer-wise relevance propagation for transformers". In: (2024). arXiv: 2402.05602.

[Arr+16] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "Explaining Predictions of Non-Linear Classifiers in NLP". In: *Proceedings of the 1st Workshop on Representation Learning for NLP*. Ed. by Phil Blunsom, Kyunghyun Cho, Shay Cohen, et al. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1–7.

[ASP19] Yannis Assael, Thea Sommerschield, and Jonathan Prag. "Restoring ancient text using deep learning: a case study on Greek epigraphy". 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 Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6368–6375.

[Ass+22] Yannis Assael, Thea Sommerschield, Brendan Shillingford, et al. "Restoring and attributing ancient texts using deep neural networks". In: *Nature* 603.7900 (2022), pp. 280–283.

[Ass+25] Yannis Assael, Thea Sommerschield, Alison Cooley, et al. "Contextualizing ancient texts with generative neural networks". In: *Nature* (2025), pp. 1–7.

[Bac+15] Sebastian Bach, Alexander Binder, Grégoire Montavon, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

[Bar+20] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

[BB20] David Bamman and Patrick J. Burns. *Latin BERT: A Contextual Language Model for Classical Philology*. 2020. arXiv: 2009.10053 [cs.CL].

[Bro+20] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[BS24] Jaydeep Borkar and David A. Smith. *Mind the Gap: Analyzing Lacunae with Transformer-Based Transcription*. 2024. arXiv: 2407.00250 [cs.CV].

[CLL19] Hugh Chen, Scott Lundberg, and Su-In Lee. *Explaining Models by Propagating Shapley Values of Local Components*. 2019. arXiv: 1911.11888 [cs.LG].

[Deh25] Tahereh Dehdarirad. "Evaluating explainability in language classification models: A unified framework incorporating feature attribution methods and key factors affecting faithfulness". In: *Data and Information Management* (2025), p. 100101.

[Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[DeY+20] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, et al. "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 4443–4458.

[DK17] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].

[Dob21] James Dobson. "Interpretable Outputs: Criteria for Machine Learning in the Humanities." In: *DHQ: Digital Humanities Quarterly* 15.2 (2021).

[Eng23] Joseph Enguehard. "Sequential integrated gradients: a simple but effective method for explaining language models". In: *arXiv:2305.15853* (2023).

[Fet+20] Ethan Fetaya, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. "Restoration of fragmentary Babylonian texts using recurrent neural networks". In: *Proceedings of the National Academy of Sciences* 117.37 (2020), pp. 22743–22751. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2003794117.

[Gat25] Gabriele Gattiglia. "Managing Artificial Intelligence in Archeology. An overview". In: *Journal of Cultural Heritage* 71 (2025), pp. 225–233.

[Has+24] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, et al. "Interpreting black-box models: a review on explainable artificial intelligence". In: *Cognitive Computation* 16.1 (2024), pp. 45–74.

[HLL25] Sungmin Han, Jeonghyun Lee, and Sangkyun Lee. "Contrast-CAT: Contrasting Activations for Enhanced Interpretability in Transformer-based Text Classifiers". In: *The 41st Conference on Uncertainty in Artificial Intelligence*. 2025.

[HYL17] Will Hamilton, Zhitao Ying, and Jure Leskovec. "Inductive Representation Learning on Large Graphs". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc., 2017.

[Jac+21] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, et al. "Contrastive Explanations for Model Interpretability". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1597–1611.

[JG20] Alon Jacovi and Yoav Goldberg. "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 4198–4205.

[JGS20] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. "Drug discovery with explainable artificial intelligence". In: *Nature Machine Intelligence* 2.10 (2020), pp. 573–584.

[JK02] Kalervo Järvelin and Jaana Kekäläinen. "Cumulated gain-based evaluation of IR techniques". In: *ACM Transactions on Information Systems (TOIS)* 20.4 (2002), pp. 422–446.

[JSS19] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. "What Does BERT Learn about the Structure of Language?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3651–3657.

[Kan+21] Kyeongpil Kang, Kyohoon Jin, Soyoung Yang, et al. "Restoring and Mining the Records of the Joseon Dynasty via Neural Language Modeling and Machine Translation". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, et al. Online: Association for Computational Linguistics, June 2021, pp. 4031–4042.

[Laz+21] Koren Lazar, Benny Saret, Asaf Yehudai, et al. "Filling the Gaps in Ancient Akkadian Texts: A Masked Language Modelling Approach". In: Jan. 2021, pp. 4682–4691.

[Li+22] Xuhong Li, Haoyi Xiong, Xingjian Li, et al. "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond". In: *Knowledge and Information Systems* 64.12 (2022), pp. 3197–3234.

[Li+23] Dongfang Li, Zetian Sun, Xinshuo Hu, et al. *A Survey of Large Language Models Attribution*. 2023. arXiv: 2311.03731 [cs.CL].

[Lin+22] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. "A survey of transformers". In: *AI open* 3 (2022), pp. 111–132.

[LL17] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI].

[Mol25] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 3rd ed. Lulu. com, 2025.

[Mon+19] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. "Layer-wise relevance propagation: an overview". In: *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), pp. 193–209.

[Mün+24] Sander Münster, Ferdinand Maiwald, Isabella di Lenardo, et al. "Artificial Intelligence for Digital Heritage Innovation: Setting up a R&D Agenda for Europe". In: *Heritage* 7.2 (2024), pp. 794–816.

[Par19] R Parker. "Data in online database 'Lexicon of Greek Personal Names (LGPN)'". In: (2019).

[PKO23] Katerina Papavassileiou, Dimitrios I. Kosmopoulos, and Gareth Owens. "A Generative Model for the Mycenaean Linear B Script and Its Application in Infilling Text from Ancient Tablets". In: *J. Comput. Cult. Herit.* 16.3 (Aug. 2023).

[RKR21] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. "A primer in BERTology: What we know about how BERT works". In: *Transactions of the association for computational linguistics* (2021), pp. 842–866.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016. arXiv: 1602.04938 [cs.LG].

[Sal94] Benet Salway. "What's in a Name? A Survey of Roman Onomastic Practice from c. 700 BC to AD 700". In: *The Journal of Roman Studies* 84 (1994), pp. 124–145.

[Sam+21] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. "Explaining deep neural networks and beyond: A review of methods and applications". In: *Proceedings of the IEEE* 109.3 (2021), pp. 247–278.

[Sel+19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359.

[She+20] Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. "Blank Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 5186–5198.

[Shi22] Adam Shimi. *Interpretability: Integrated Gradients is a decent feature importance measure*. https://www.lesswrong.com/posts/Rv6ba3CMhZGZzNH7x/interpretability-integrated-gradients-is-a-decent. Accessed: 2024-06-27. 2022.

[Shr+16] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. "Not just a black box: Learning important features through propagating activation differences". In: *arXiv:1605.01713* (2016).

[Som+23] Thea Sommerschield, Yannis Assael, John Pavlopoulos, et al. "Machine Learning for Ancient Languages: A Survey". In: *Computational Linguistics* 49.3 (Sept. 2023), pp. 703–747. eprint: https://direct.mit.edu/coli/article-pdf/49/3/703/2177413/coli_a_00481.pdf.

[SR21] Soumya Sanyal and Xiang Ren. *Discretized Integrated Gradients for Explaining Language Models*. 2021. arXiv: 2108.13654 [cs.CL].

[SSN24] Priscylla Silva, Claudio T. Silva, and Luis Gustavo Nonato. *Exploring the Relationship Between Feature Attribution Methods and Model Performance*. 2024. arXiv: 2405.13957 [cs.LG].

[STY17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 3319–3328.

[SVZ13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv:1312.6034* (2013).

[SWM17] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. 2017. arXiv: 1708.08296 [cs.AI].

[Tay23] Mohammad Mustafa Taye. "Understanding of machine learning with deep learning: architectures, workflow, applications and future directions". In: *Computers* 12.5 (2023), p. 91.

[TDP19] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "BERT Rediscovers the Classical NLP Pipeline". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4593–4601.

[Val+23] Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, et al. "The geometry of hidden representations of large transformer models". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, et al. Vol. 36. Curran Associates, Inc., 2023, pp. 51234–51252.

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc., 2017.

[VEA22] Daniel Vale, Ali El-Sharif, and Muhammed Ali. "Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law". In: *AI and Ethics* 2.4 (2022), pp. 815–826.

[Voo+99] Ellen M Voorhees et al. "The trec-8 question answering track report." In: *Trec*. Vol. 99. 1999, pp. 77–82.

[Wan+23] Dongbo Wang, Chang Liu, Zhao Zhixiao, et al. *GujiBERT and GujiGPT: Construction of Intelligent Information Processing Foundation Language Models for Ancient Texts*. July 2023.

[WO21] Zhengxuan Wu and Desmond C Ong. "On explaining your explanations of bert: An empirical study with sequence classification". In: *arXiv.2101.00196* (2021).

[ZF13] Matthew D Zeiler and Rob Fergus. *Visualizing and Understanding Convolutional Networks*. 2013. arXiv: 1311.2901 [cs.CV].

List of Acronyms

AUEB Athens University of Economics and Business

AI Artificial Intelligence

ML Machine Learning

NLP Natural Language Processing

XAI Explainable Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

GPT Generative Pretrained Transformer

RNN Recurrent Neural Network

ResNet Residual Neural Network

MHSA Multi-Head Self-Attention

MHA Multi-Head Attention

FFN Feed-Forward Network

PE Positional Encoding

PHI Packard Humanities Institute

I.PHI PHI Greek Inscriptions Corpus

LGPN Lexicon of Greek Personal Names

LED Latin Epigraphic Dataset

EDR Epigraphic Database Roma

EDH Epigraphic Database Heidelberg

EDCS Epigraphik-Datenbank Clauss–Slaby

BCE Before the Common Era

CE Common Era

DH Digital Humanities

IG Integrated Gradients

SeqIG Sequential Integrated Gradients

LRP Layer-wise Relevance Propagation

AttnLRP Attention-Aware Layer-wise Relevance Propagation

LIME Local Interpretable Model-agnostic Explanations

SHAP SHapley Additive exPlanations

OCR Optical Character Recognition

MRR Mean Reciprocal Rank

MAP Mean Average Precision

DCG Discounted Cumulative Gain

IDCG Ideal Discounted Cumulative Gain

nDCG Normalized Discounted Cumulative Gain

PR Precision–Recall

ROC Receiver Operating Characteristic

AUROC Area Under the ROC Curve

AUPRC Area Under the Precision–Recall Curve

TP True Positive

FP False Positive

TN True Negative

FN False Negative

F1 F1 Score

L2 Euclidean Norm (L2)

WLS Weighted Least Squares

UNK Unknown Token

CLS Classification Token

BPE Byte-Pair Encoding

HITL Human-in-the-Loop

List of Figures

List of Tables