DART-FT: Decomposed, Ambiguity-Aware, Risk-Targeted Fine-Tuning of Transformer Models for Ancient Greek Morphological Tagging

by Matthew K. Weber

B.A. in Computer Science, December 2001,
Virginia Polytechnic Institute and State University
M.A. in Practical Theology, December 2009, Regent University

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

August 31, 2026

Praxis directed by

Vijay Raghavan
Professorial Lecturer of Engineering and Applied Science

Approved for Public Release; Distribution Unlimited.
Public Release Case Number 26-1036

The School of Engineering and Applied Science of The George Washington University certifies that Matthew K. Weber has passed the Final Examination for the degree of Doctor of Engineering as of July 9, 2026. This is the final and approved form of the Praxis.

DART-FT: Decomposed, Ambiguity-Aware, Risk-Targeted Fine-Tuning of Transformer Models for Ancient Greek Morphological Tagging

Matthew K. Weber

Praxis Research Committee:

Vijay Raghavan, Professorial Lecturer of Engineering and Applied Science, Praxis Director

Amir Etemadi, Adjunct Professor of Engineering and Applied Science, Committee Chair

Mazen Mheish, Professorial Lecturer of Engineering and Applied Science, Committee Member

© Copyright 2026 The MITRE Corporation.
ALL RIGHTS RESERVED.

Dedication

To my children Emma, Joshua, and Kate, thank you for your encouragement through many days and nights of study and research. I believe Artificial Intelligence is poised to open possibilities for your lives well beyond what we can currently imagine. My study in this topic is largely for you, to provide you insights that you can leverage to change the world for the better. To my amazing wife Su Chong: your patience, and encouragement throughout my studies repeatedly helped me move forward, especially in the most difficult times. Mom, Dad, 어머니, 아버지, Tim, Andrew, and Rachel, thank you for your constant support and encouragement.

Acknowledgements

The author would like to acknowledge and express his gratitude and appreciation for the consistent and thoughtful advising provided by Dr. Vijay Raghavan. His support and guidance throughout the process of developing this praxis were invaluable and strengthened this work.

The author also wishes to thank his employer, The MITRE Corporation, for the support and emphasis on continuous learning, which helped enable this work and its potential future applications.

The author's affiliation with The MITRE Corporation is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author. ©2026 The MITRE Corporation. ALL RIGHTS RESERVED.

Abstract of Praxis

DART-FT: Decomposed, Ambiguity-Aware, Risk-Targeted Fine-Tuning of Transformer Models for Ancient Greek Morphological Tagging

Low-resource languages present challenges for generating accurate morphological tagging models due to data sparsity. Languages which exhibit complex morphological structures (internal word structures) further compound the problem of accurate morphological tagging model creation. Ancient Greek is a low-resource language with complex morphology, leading to frequent morphological ambiguity where a single word form can have multiple grammatical interpretations. Standard fine-tuning techniques, such as training with cross-entropy loss, focus on increasing overall accuracy, but can leave systematic errors concentrated in the most ambiguous tokens. Decomposed, Ambiguity-Aware, Risk-Targeted Fine-Tuning (DART-FT) is a targeted fine-tuning method that identifies a slice of tokens (the high-risk slice) associated with recurrent baseline model errors. The method applies multiple filtering passes to dynamically determine which tokens to target for fine-tuning. The targeted fine-tuning process utilizes oversampling. Experiments were conducted with cross-entropy, class-balanced, and class-balanced combined with focal loss, producing a total of 90 distinct models over 5 random seeds on the UD Ancient Greek PROIEL treebank. The model with the largest increase in accuracy on the high-risk slice, as compared to the corresponding standard cross-entropy fine-tuned baseline model, exhibited a statistically significant increase of 6.4 percentage points. Global macro F1 scores for the models were stable and slightly increased. DART-FT enables targeted repairs to error-prone areas of fine-grained

morphological tagging models trained on limited labeled corpora and serves as a resource-efficient alternative when additional pre-training is impractical.

Table of Contents

Dedication iv

Acknowledgements v

Abstract of Praxis vi

List of Figures xiii

List of Tables xiv

List of Equations xvi

List of Symbols xvii

List of Acronyms xviii

Glossary of Terms xx

Chapter 1—Introduction 1

1.1 Background 1

1.2 Research Motivation 4

1.3 Problem Statement 5

1.4 Thesis Statement 5

1.5 Research Objectives 6

1.6 Research Questions and Hypotheses 6

1.7 Scope of Research 7

1.8 Research Limitations 7

1.9 Organization of Praxis 8

Chapter 2—Literature Review 10

2.1 Introduction 10

2.2 Ancient Greek Morphology 10

2.3 Corpus and Treebanks 11
2.4 Tagging and Granularity 12
2.5 Morphological Ambiguity 13
2.6 Impacts of Morphological Ambiguity 14
2.7 Morphological Error Analysis 15
2.8 Error Propagation in Downstream Tasks 16
2.9 Architectures for Part-of-Speech and Morphological Tagging 16
2.10 Fine-Tuning Approaches 18
2.10.1 Loss-based Approaches 19
2.10.2 Data-driven Approaches 21
2.10.3 Architecture-based Approaches 22
2.10.4 Iterative Error-Driven Approaches 22
2.11 Summary and Conclusion 23
Chapter 3—Methodology 25
3.1 Introduction 25
3.2 Experimental Pipeline Overview 25
3.3 Data and Data Preparation 27
3.3.1 Data Normalization 27
3.3.2 Tokenization 28
3.3.3 Morphological Feature Representation 28
3.4 Identification of the High-Risk Slice 28
3.4.1 Error Driven Mining (Pass 1) 30
3.4.2 Imbalanced Feature-Value Pair Filtering (Pass 2) 31

3.4.3 Slice Freezing and Mask Creation (Pass 3) 34
3.5 Fine-tuning Methodology 34
3.5.1 Baseline Standard Fine-Tuning 35
3.5.2 Targeted Fine-Tuning 36
3.6 Evaluation Metrics and Statistical Testing 38
3.6.1 Evaluation Metrics 38
3.6.2 Statistical Testing 40
3.6.3 Multi-seed 44
3.7 Implementation Details 44
3.8 Downstream Translation Pilot (PROIEL Ancient Greek to Latin) 44
Chapter 4—Results 46
4.1 Introduction 46
4.2 High-Risk Slice Characteristics 47
4.2.1 Filtering Process Results (Passes 1-3) 47
4.2.2 Slice Composition 49
4.2.3 Token Level Composition 54
4.3 High-Risk Slice Performance 55
4.3.1 Evaluation Setup 55
4.3.2 Quantitative Performance Improvements 56
4.3.3 Error Correction Analysis 58
4.3.4 Statistical Significance Analysis 59
4.3.5 Summary of Support for Hypothesis 2 60
4.4 Sensitivity Analysis 60

4.4.1 Oversampling Factor Impact Analysis 60
4.4.2 Impact of Loss Function Hyperparameters 62
4.4.3 Threshold Ablation Study 65
4.4.4 Summary of Sensitivity Analysis and Support for Hypothesis 3 68
4.5 Impact on all Slices 68
4.6 Error Topology and Class Confusion Analysis 69
4.6.1 Systematic Confusion Reduction (Gender) 71
4.6.2 Systematic Confusion Reduction (Tense) 72
4.6.3 Confusion Regression (Case) 73
4.6.4 Feature Stability 74
4.7 Training Dynamics 74
4.8 Downstream Translation Results 77
4.9 Summary 77
Chapter 5—Discussion and Conclusions 79
5.1 Discussion 79
5.1.1 Error Reduction in Gender Feature Syncretism 79
5.1.2 CE and Class Balanced Loss Robustness vs. Peak Performance 81
5.1.3 Multi-label Error Regression in Case and PronType Features 82
5.1.4 Global Stability and Catastrophic Forgetting Avoidance 83
5.1.5 Applicability and Generalizability Boundaries 84
5.1.6 Translation Pilot Implications 85
5.2 Conclusions 85
5.2.1 Research Question 1 Evaluation and Conclusion 85

5.2.2 Research Question 2 Evaluation and Conclusion 86

5.2.3 Research Question 3 Evaluation and Conclusion 86

5.3 Contributions to the Body of Knowledge 87

5.4 Recommendations for Future Research 88

References 90

Appendix A 106

Ancient Greek Language Time Periods and Evolution 109

List of Figures

Figure 1 Image of John 1:1-3 from the Codex Sinaiticus Project..... 2

Figure 2 PROIEL Universal Dependency Ancient Greek Treebank Structure. 12

Figure 3 Pipeline depicting the end-to-end targeted fine-tuning 26

Figure 4 Overview of the methodology to identify the high-risk slice..... 30

Figure 5 Feature category inclusion in the high-risk slice..... 51

Figure 6 Feature-value pair inclusion in the high-risk slice 52

Figure 7 Feature-value pair composition of the high-risk slice 54

Figure 8 Frequency of the most common Ancient Greek tokens 55

Figure 9 Fine-grained morphological accuracy comparison 57

Figure 10 Errors involving Gender=Masc,Neut 59

Figure 11 Model performance for various oversampling factors 63

Figure 12 Delta confusion matrices for morphological features. 70

Figure 13 Gender feature delta confusion matrix 72

Figure 14 Tense features delta confusion matrix..... 73

Figure 15 Case feature delta confusion matrix 74

Figure 16 Comparison of training loss across epochs 75

Figure 17 Model Initialization Flow. 76

Figure 18 Raw execution logs of the high-risk slice identification pipeline 107

Figure 19 Pseudo code algorithm of the DART-FT methodology 108

Figure 20 Experimental pipeline details 109

List of Tables

Table 1 Example of τίνα ambiguous forms 4

Table 2 Illustrative examples of Ancient Greek words 14

Table 3 Ancient Greek model accuracy metrics summary 18

Table 4 UD Ancient Greek PROIEL dataset composition statistics 27

Table 5 High-risk slice candidate identification example 31

Table 6 Constant Parameters used during the training 35

Table 7 Variable Parameters used during the training and evaluation 35

Table 8 McNemar 2x2 matrix 41

Table 9 Threshold parameters utilized in creation of the high-risk slice 47

Table 10 High-Risk Slice development for each seed through each pass 49

Table 11 Summary metrics for each pass averaged across the five seeds 49

Table 12 Number of feature categories included in the high-risk slice 50

Table 13 Comparison of baseline fine-tuned and targeted fine-tuned models 57

Table 14 Illustrative examples of token fine-grained morphology corrected 58

Table 15 Average accuracy changes on the high-risk slice 61

Table 16 Oversampling highest performing models accuracy change 61

Table 17 Average accuracy changes on the high-risk slice 64

Table 18 Accuracy change on the high-risk slice for the highest performing 65

Table 19 Threshold Ablation Slice Composition Results 67

Table 20 Threshold Ablation Performance Results 67

Table 21 Threshold Ablation Multi-Seed Diversity Cap Mean Results 68

Table 22 Global F1 change for the highest performing models for each seed 69

Table 23 Validation Accuracy Change per Epoch..... 77

Table 24 High-risk feature-value pairs for each of the 5 random seeds 108

Table 25 Time Periods of Ancient Greek Language 110

List of Equations

Equation 1 Relative improvement with accurate morphology 15

Equation 2 Relative reduction of baseline 15

Equation 3 Cross Entropy Loss Function 20

Equation 4 Focal Loss Function 20

Equation 5 Class-Balanced Focal Loss Function..... 21

Equation 6 Formal definition of the high-risk slice candidates (pass 1)..... 31

Equation 7 High-risk slice is frozen based on the filtering candidates (pass 3) 34

Equation 8 Class Balanced Weight function 37

Equation 9 Formal definition of F1 score metric..... 39

Equation 10 Macro F1 score definition 39

Equation 11 Exact Two-sided McNemar Test..... 42

Equation 12 Agresti-Min Wald+2 Confidence Interval Standard Error..... 43

Equation 13 Agresti-Min Wald+2 95% Confidence Interval 43

List of Symbols

maF_1	Macro-averaged F_1 score
τ_e	Error rate threshold for high-risk slice identification
τ_f	Minimum frequency threshold for high-risk slice identification
n_{12}	Count of instances where baseline is correct and DART-FT is incorrect
n_{21}	Count of instances where baseline is incorrect and DART-FT is correct
β	Smoothing hyperparameter for Class Balanced loss
γ	Focusing parameter for Focal loss
w_c	Class-balanced weighting factor for class c

List of Acronyms

AGDT	Ancient Greek Dependency Treebank
ALP	Ancient Language Processing
BCE	Before Common Era
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BLEU	Bilingual Evaluation Understudy
CE	Cross-Entropy
CI	Confidence Interval
CNN	Convolutional Neural Network
CoNLL-U	Conference on Computational Natural Language Learning – Universal
DART-FT	Decomposed, Ambiguity-Aware, Risk-Targeted Fine-Tuning
CRF	Conditional Random Field
FV	Feature-Value
HMM	Hidden Markov Model
HRL	High-Resource Language
JTT	Just Train Twice
LRL	Low-Resource Language
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
OS	Oversampling
POS	Part-of-Speech
PROIEL	Pragmatic Resources in Old Indo-European Languages

RO	Research Objective
RQ	Research Question
SE	Standard Error
SMOTE	Synthetic Minority Over-sampling Technique
UD	Universal Dependencies
UFEATS	Universal Morphological Features
UPOS	Universal Part-of-Speech
TER	Translation Edit Rate

Glossary of Terms

Baseline Model: Standard fine-tuned Ancient-Greek-BERT, trained with cross-entropy loss and no oversampling

Feature-Value Pair (FV): A single unit of morphological information which includes an attribute and associated grammatical value

High-Risk Slice: An empirically identified subset of tokens in the dataset where the baseline model exhibits frequent and systematic fine-grained morphological tagging errors, defined through a three-pass feature-value filtering procedure

Low-Resource Language: A language with relatively minimal digital data or natural language processing tools

Morphological Ambiguity: A word which can have multiple syntactic meanings depending on the grammatical interpretation

Morphology: The internal grammatical structure of a word (e.g., case, gender, number)

Oversampling: Data augmentation that increases the learning process exposure to a subset of data

Random Seed: An initial value for a random number generator

Scriptio Continua: Text written without spaces and sometimes also lacking punctuation or diacritics

Syntactic: Grammatical rules that determine how sentences are formed

Syncretism: The exact same word utilizing different grammatical categories to express different meanings which is resolved by context

Treebank: A corpus which is annotated with morphological labels and structure

Chapter 1—Introduction

1.1 Background

Accurate natural language processing is a foundational activity which underpins *essential* aspects of modern life in areas such as international commerce, cross-cultural relationships, and academic exploration. Automated language processing tools exist and are increasing in sophistication for high resource languages (HRLs) such as English, French, and Spanish. However, a divide exists when compared to low resource language (LRL) environments. For example, human translators remain the primary agent performing translation activities, resulting in a process which is slow, expensive, and difficult to scale. Although different definitions exist for what constitutes a LRL (Nigatu et al., 2024), they typically involve a lack of digital resources, linguistic data, and natural language processing tools (Pakray et al., 2025). Morphological ambiguity, where a word or character can hold multiple grammatical meanings (Prior et al., 2010), is often a limitation preventing LRL automated processing. Addressing the morphological ambiguity challenge helps to close the HRL/LRL divide, enabling efficient flow of information which could accelerate academic discovery and natural language processing activities (Pava et al., 2025).

To enable automated LRL processing, there is a need for some fundamental underpinnings to be in place. For example, each word in a language requires descriptive metadata labels. These labels can include noun/verb/etc. and are frequently stored in a data structure known as a treebank. The labeling of each word (and dependencies between words) in the language is a manual and time-intensive process carried out by linguistic experts. The accurate labeling of each word can support automated translation

and other downstream tasks. In some languages, certain words have a higher potential for label errors than other words. The words with high label error rates can cause negative impacts to translation.

One such LRL is Ancient Greek. Ancient Greek contains a number of linguistic features which lead to ambiguity in general, such as a lack of punctuation and spaces between words in written ancient manuscripts (i.e., *Scriptio Continua*) (Heilmann, 2020, p. 181) as noted in Figure 1. More important as relates to this praxis, the morphology of Ancient Greek includes features such as "part of speech, person, number, tense, mood, voice, gender, case and degree" where the combination of different feature expressions can lead to the exact same word having different meanings in different contexts (Bamman & Crane, 2011, p. 83).

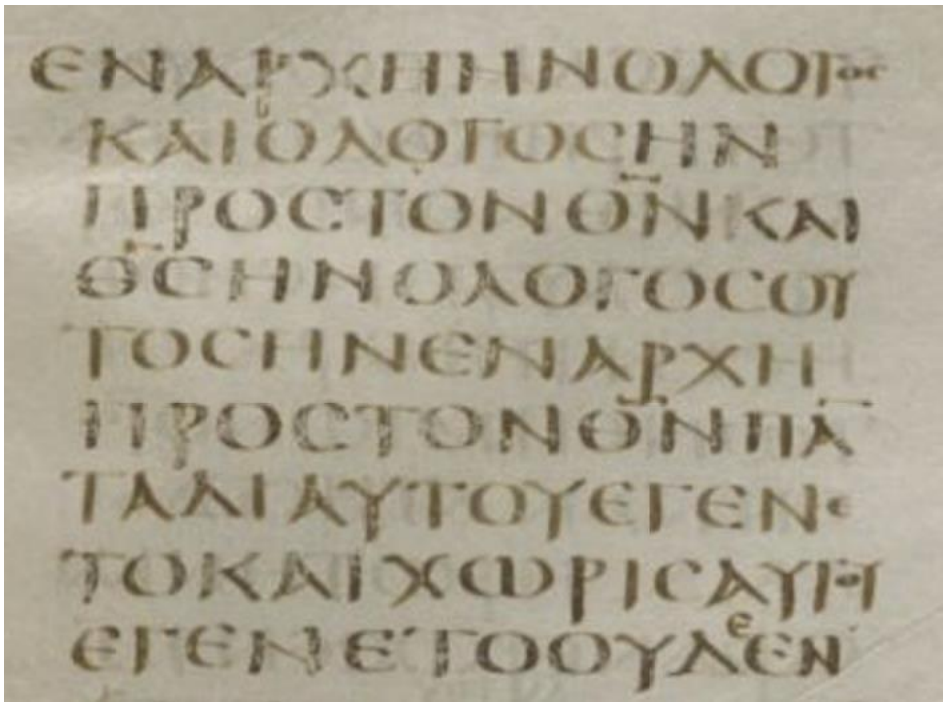

Figure 1 Image of John 1:1-3 from the Codex Sinaiticus Project
(Codex Sinaiticus - See The Manuscript | John 1:1-3, Folio 247, British Library, 2025) Non-commercial educational use permitted.

A concrete example of morphological ambiguity (specifically syncretism) is helpful in conveying the concept. The word τίς is an interrogative pronoun (*Strong's Greek: 5101. Τίς*, 1890) and can take several forms depending on the use within a sentence, one of which is τίνα (Table 1). The English translation of τιν-ά in the plural nominative form is *what things*[1] and the plural accusative form is *what/which things.* [2] Additionally, as a general source of ambiguity, Ancient Greek lacks accent marks in many manuscripts, which can result in the same written word having different meanings depending on context. The lack of accent marks can contribute to broader ambiguities. However, this praxis focuses on the morphological ambiguity which can arise from syncretism and data sparsity. Language models, such as Ancient-Greek-BERT, are challenged by morphological ambiguity, often leading to labeling errors. This praxis proposes a novel method to identify an empirically defined high-risk slice of Ancient Greek tokens associated with frequent and systematic fine-grained morphological tagging errors and evaluates the impact of a targeted fine-tuned Ancient-Greek-BERT model in reducing labeling errors for those tokens.

[1] John 10:6 τίνα takes the plural nominative form – "his parable spake Jesus unto them: but they understood not *what things* they were which he spake unto them. " KJV https://biblehub.com/interlinear/john/10.htm; https://biblehub.com/interlinear/john/10-6.htm

[2] Acts 17:20 τίνα takes the plural accusative form - "For thou bringest certain strange things to our ears: we would know therefore *what these things* mean." KJV https://biblehub.com/kjvs/acts/17.htm; https://biblehub.com/interlinear/acts/17.htm

Table 1 Example of τίνα ambiguous forms

Case	Interrogative			Indefinite		
	Masculine/ Feminine	**Gender-invariant**	**Neuter**	**Masculine/ Feminine**	**Gender-invariant**	**Neuter**
SINGULAR						
Nominative	τίς		τί	τὶς		τὶ
Genitive		τίν-ος, τοῦ			τιν-ός, τοῦ	
Dative		τίν-ι, τῷ			τιν-ί, τῷ	
Accusative	τίν-α		τί	τινά		τὶ
PLURAL						
Nominative	τίν-ες		τίν-α	τιν-ές		τιν-ά
Genitive		τίν-ων			τιν-ῶν	
Dative		τί-σι（ν）			τι-σί（ν）	
Accusative	τίν-ας		τίν-α	τιν-άς		τιν-ά

Note. Adapted from (Smyth, 1920, Sections 334–335).

1.2 Research Motivation

The research community recognizes the problem of labeling errors in Ancient Greek and other similar languages. Recent conferences and calls for papers have encouraged further research in this area. One such example is the 2025 Ancient Languages Processing (ALP) workshop (*ALP 2025*, 2025), where the call for papers included morphological analysis (*The Second Workshop on Ancient Language Processing @NAACL 2025*, 2025) and the papers received included several related to tagging and disambiguation (Chiruzzo et al., 2025).

A secondary motivation is to contribute towards reducing the time required for academic and applied work with Ancient Greek texts. A specific example of ongoing work with Ancient Greek is Bible translation. The Bible continues to be translated from Ancient Greek (and Hebrew/Aramaic) into modern LRLs. Over 2,500 languages are either awaiting Bible translation to begin or currently in the initial stages of the

translation process ("2024 Global Scripture Access," 2025)[3]. Improved label accuracy within an Ancient Greek language model could potentially increase the efficiency of translation, parsing, and other downstream activities. However, the primary contribution is improved fine-grained morphological tagging accuracy, not an evaluation of potential impacts on translation speed or cost.

1.3 Problem Statement

Automatic fine-grained tagging mislabels approximately 7% of Ancient Greek words, in part due to grammatical ambiguity (Singh et al., 2021, p. 134), and contributes to downstream issues such as a 26% relative reduction in machine translation quality compared with models supplied with accurate grammatical properties (Rapacz & Smywiński-Pohl, 2025, p. 145).

Several downstream issues occur because of inaccurate morphological tagging models. For example, some Ancient-Greek translation systems utilize lower-level mechanisms, such as parsers which structure syntactic information. The accuracy of the information contained within parser output positively benefits from accurate morphological label inputs. Further, suboptimal parsers limit the ability of scholars to search and compare texts (Bamman & Crane, 2011, p. 15). Although errors occur with automated tagging processes in both coarse and fine-grained tagging, this praxis focuses on fine-grained tagging as it is the more challenging problem area.

1.4 Thesis Statement

A targeted fine-tuned Ancient-Greek-BERT model will reduce mislabeling in fine-grained Ancient Greek morphological tagging.

[3] One of the larger Bible translation organizations is Wycliffe®, with annual revenues exceeding $200 million (*Wycliffe Financial Accountability*, 2025).

Fine-grained morphological tagging accuracy will improve through a targeted fine-tuning method that focuses on an empirically identified high-risk slice of tokens associated with frequent and systematic fine-grained morphological tagging errors. This research evaluates if such a targeted intervention can improve performance on error prone slices without harming global model stability.

1.5 Research Objectives

The following are the objectives:

RO1: Develop and execute a methodology for identifying an empirically defined high-risk slice of Ancient Greek tokens associated with frequent and systematic morphological tagging errors.

RO2: Develop a targeted fine-tuned Ancient-Greek-BERT model with improved fine-grained morphological tagging accuracy compared to baseline standard fine-tuned Ancient-Greek-BERT models, for the high-risk slice.

RO3: Provide statistical validation and error analysis of the targeted fine-tuned model compared to baseline standard fine-tuned models.

1.6 Research Questions and Hypotheses

The following research questions will support the research objectives:

RQ1: Does targeted fine-tuning negatively impact global macro F1 performance compared with a standard fine-tuned model?

RQ2: Does targeted fine-tuning increase fine-grained morphological tagging accuracy for the high-risk slice compared to a standard fine-tuned Ancient-Greek-BERT model?

RQ3: How sensitive is targeted fine-tuning performance to the degree of oversampling of high-risk tokens?

H1: Targeted fine-tuning does not negatively impact global macro F1 performance compared with a standard fine-tuned model.

H2: Targeted fine-tuning yields significantly higher fine-grained morphological tagging accuracy for the high-risk slice compared to a standard fine-tuned Ancient-Greek-BERT model.

H3: High-risk slice performance is sensitive to the oversampling factor, with intermediate oversampling achieving the highest mean accuracy improvement among tested factors.

1.7 Scope of Research

This praxis focuses on fine-grained morphological tagging (as compared to coarse grained morphological tagging). The research utilizes the Ancient-Greek-BERT model using the PROIEL Universal Dependency Treebank. Further, a small-scale exploratory downstream translation pilot to evaluate if the application of DART-FT can impact translation quality (however, downstream translation quality is a secondary evaluation compared with the primary tagging work) is included.

1.8 Research Limitations

A few areas of research limitations exist.

1. There is a limited corpus of training data, focused on the collection of 27 books of the Ancient Greek New Testament of the Bible and Herodotus. A total of approximately 140,000 tokens exist in the Ancient Greek New Testament portion of the PROIEL treebank (The PROIEL Treebank,

2014/2025). This limited and narrowly scoped training set is quite small compared with the datasets used by modern transformer-based models which can exceed billions of tokens (e.g., Colossal Clean Crawl Corpus) (Raffel et al., 2020). Further, the focus is on the PROIEL treebank and does not consider other datasets such as MorphGNT (*MorphGNT*, 2025).

2. Although generalizability of the methodology is plausible, it is not empirically validated. Applicability of the method depends on several prerequisites to include: a labeled corpus (e.g., treebank) with token-level gold annotations, a decomposable finite set of feature-value pairs, and a pretrained transformer model which can be fine-tuned.
3. The definition of the high-risk slice is established within this research, suggesting that other definitions may yield different results.
4. A translation pilot is presented as a secondary exploratory result.
5. The threshold values (e.g., 35% diversity limit) used to identify the high-risk slice were selected as design choices and significantly impact the slice composition. A limited ablation study utilizing alternative values for the threshold parameters was conducted for the purpose of identifying sensitivity. However, an exhaustive optimization study was outside the scope of this praxis.

1.9 Organization of Praxis

This praxis is structured in the following manner: Chapter 2 reviews the relevant literature related to Ancient Greek natural language processing and fine-tuning methods. Chapter 3 elaborates on the methodology of targeted-fine tuning and evaluation. Chapter

4 presents the results to include statistical validation, demonstrating support for the hypotheses. Lastly, Chapter 5 provides a discussion of the conclusions as well as potential areas for future research.

Chapter 2—Literature Review

2.1 Introduction

The literature related to this topic area is expansive, in some cases covering hundreds of years of history (specifically, the study of the Ancient Greek language[4]). This literature review will include a high-level overview of relevant items of interest related to Ancient Greek morphology. A summary of the characteristics of Ancient Greek that lead to morphological ambiguity will be discussed. The specific corpus and treebank will be explored. Then the areas of modern natural language processing, especially as relates to fine-tuning methods, and evaluation considerations will be reviewed.

2.2 Ancient Greek Morphology

The study of the internal structure of words is the field of morphology (J. T. Devlin et al., 2004, p. 14984). Morphemes are the components which form words. For example, morphemes such as a prefix + stem (e.g., *un* + *happy*) could be combined to form a word (e.g., *unhappy*). Ancient Greek morphology contains several forms which differ from English[5]. For example, nouns inflect in five cases – nominative, genitive, dative, accusative, and vocative (Smyth, 1920, §201), whereas English nouns lack such a mechanism and rely mainly on word order to indicate grammatical relationships within sentences. Ancient Greek sentence construction allows for flexible word order, as the morphology provides the primary mechanism for grammatical relations (Allan, 2014, p.

[4] Although expertise in reading and interpreting Ancient Greek is helpful, it is not required to gain a robust understanding of the literature.

[5] It is beyond the scope of this praxis to detail each of the morphological features of Ancient Greek. However, several public domain resources are available to provide additional insight.

181). Therefore, accurate morphological features are essential to derive accurate syntactic meaning in Ancient Greek.

2.3 Corpus and Treebanks

In natural language processing (NLP), "a treebank is a large collection of sentences in which the syntactic relation for every word is made explicit – where a human has encoded an interpretation of the sentence in the form of a linguistic annotation." (Bamman & Crane, 2011, p. 80) Treebanks typically involve human expert annotation (or review of automated annotation) of morphology for each word in the corpora. These annotations can then be used by automated taggers and parsers for training and evaluation (Marcus et al., 1993, p. 313).

The Universal Dependencies (UD) Framework was developed to provide a standard structure to express annotations across treebanks and enables a level of comparison between treebanks. Three significant elements of the UD include Universal Part-of-Speech (UPOS), Universal Morphological Features (FEATS), and dependency relation labels (de Marneffe et al., 2021, pp. 258, 261, 263). The elements are typically encoded through the Conference on Computational Natural Language Learning (CoNLL-U) file format (Zeman et al., 2017). The format is tab separated and contains ten columns: ID, FORM, LEMMA, UPOS, XPOS, FEATS, HEAD, DEPREL, DEPS, MISC (*CoNLL-U Format*, 2025)

The Pragmatic Resources in Old Indo-European Languages (PROIEL) project developed multiple parallel treebanks focused on old Indo-European languages (Eckhoff et al., 2018). The project produced several software tools and process guidelines. Annotation is performed manually by trained human annotators and includes editorial

review (Eckhoff et al., 2018, p. 34). The Ancient Greek PROIEL treebank was developed with the Ancient Greek New Testament and portions of Herodotus (*UD Ancient Greek PROIEL*, 2025). Because these texts originate from different periods of Ancient Greek, morphological variation exists within the treebank. Figure 2 represents an excerpt of data from the PROIEL treebank, demonstrating the structure of the UPOS single element and FEATS concatenation of elements. Other Ancient Greek treebank resources exist, but differences exist related to the annotation workflows. For example, the Ancient Greek and Latin Dependency Treebank (AGDT)/Perseus utilizes multiple annotators with adjudication (Bamman et al., 2009, p. 5) whereas Dependency Treebanks of Ancient Greek Prose utilize a single annotator (Gorman, 2020).

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Βίβλος	βίβλος	NOUN	Nb	Case=Nom; Gender=Fem; Number=Sing	0	root	_	ref=MATT_1.1
2	γενέσεως	γένεσις	NOUN	Nb	Case=Gen; Gender=Fem; Number=Sing	1	nmod	_	ref=MATT_1.1
3	Ἰησοῦ	Ἰησοῦς	PROPN	Ne	Case=Gen; Gender=Masc; Number=Sing	2	nmod	_	ref=MATT_1.1

Figure 2 PROIEL Universal Dependency Ancient Greek Treebank Structure.
Text is from the Gospel of Matthew 1:1

2.4 Tagging and Granularity

Tagging involves making an assignment of a part of speech or grammatical property to a particular word in a text (Merialdo, 1994, p. 155). Granularity is a key differentiator when describing morphological tagging. Coarse-grained morphological tagging refers to only the UPOS element, whereas fine-grained morphological tagging includes both UPOS and FEATS elements (Mueller et al., 2013, p. 322). For example, the English word *cats* has a coarse-grained UPOS tag of "NOUN" and a fine-grained UPOS+FEATS tag of "NOUN|Number=Plur". In Ancient Greek, the word *θεόν* (i.e., the

accusative singular form of the word θεός or *God*) has the following fine-grained tags: "NOUN|Case=Acc|Gender=Masc|Number=Sing" (*UD_Ancient_Greek-PROIEL Dev Split*, 2025, Line 1352). Due to the morphologically rich nature of Ancient Greek the fine-grained tag is more extensive than the English tag because it must store information such as case, gender, and number. The main emphasis of this research is on fine-grained tagging. As compared to coarse-grained morphological tagging, fine-grained tagging introduces the potential for more errors in morphologically ambiguous tokens because the model must learn more distinct morphological features.

2.5 Morphological Ambiguity

Ambiguity arises in morphology when a word can have multiple syntactic meanings depending on the context, including the POS tag as well as other morphological feature tags. (Tolegen et al., 2024, p. 1). In English, we see lexical ambiguity in examples such as with the word *book*. For example, we could *book (verb)* a flight or read a *book (noun)*. Instead of relying on fine-grained morphology to disambiguate the syntactic meaning, English utilizes word order for this purpose.

In contrast, Ancient Greek displays morphological ambiguity to a greater degree than English because word order is highly flexible and cannot be relied on to provide clarity on ambiguous forms. Table 2 provides illustrative examples of common frequently morphologically ambiguous Ancient Greek words and associated fine-grained morphological labels.

Table 2 Illustrative examples of Ancient Greek words which are frequently morphologically ambiguous. Adapted from (Smyth, 1920, p. §334, 338, 1689)

Morphologically Ambiguous Ancient Greek Word	Fine-grained morphological labels	English Gloss
ἐπὶ (*epi)*	Preposition (ADP) governing Dative	on, upon, near
	Preposition (ADP) governing Accusative	onto, against
	Preposition (ADP) governing Genitive	during, in the time of
τίς *(tis)*	Interrogative pronoun (Nom. Masc./Fem. Sing.)	who
	Interrogative pronoun (Gen. Masc./Fem. Sing.)	whose
ὅς *(hos)*	Relative pronoun (Nom. Masc. Sing.)	who
	Relative pronoun (Acc. Masc. Sing.)	whom

2.6 Impacts of Morphological Ambiguity

Morphological ambiguity produces several challenges related to machine learning. Primarily, morphological ambiguity can contribute to tagging errors involving the model learning incorrect patterns especially when combined with data sparsity. For example, if there are thousands of instances of a word in the masculine or feminine form, but very few in the neuter form, there is a potential for tagging errors to occur on the neuter form. This is the situation in morphologically rich languages, where certain forms of words can appear infrequently in the training data, resulting in sparsity of word forms between the train and test sets (Z. Liu & Prud'hommeaux, 2022). These errors propagate downstream, as accurate fine-grained morphological tags are especially important in morphologically rich languages to support downstream NLP tasks such as parsing and translation (Tsarfaty et al., 2013, p. 16).

Pertaining to translation, recent research demonstrates the impact of accurate morphological information on translation quality. Rapacz & Smywiński-Pohl

demonstrate the presence of accurate morphological information results in a BLEU score of 60.4 for interlinear translation between Ancient Greek and English, compared with a baseline score of 44.67 without accurate morphological information (Rapacz & Smywiński-Pohl, 2025, p. 151). This is an absolute gain of 15.73 BLEU points, which corresponds to a 35.2% relative improvement over the baseline as expressed in Equation 1.

$$\frac{60.40 - 44.67}{44.67} = 0.352\ (35.2\%)$$

Equation 1 Relative improvement with accurate morphology compared to baseline

Conversely, when there is a lack of accurate morphological information, it can be said that the baseline performs 26% lower than a morphologically aware model as expressed in Equation 2.

$$\frac{60.40 - 44.67}{60.40} = 0.26\ (26\%)$$

Equation 2 Relative reduction of baseline compared to a model with accurate morphological information

Therefore, when accurate morphology is absent, there is a substantial translation quality reduction as compared to a morphologically aware model.

2.7 Morphological Error Analysis

Analysis of morphological errors that focuses only on aggregate analysis (e.g., overall accuracy) may be insufficient to provide visibility into specific patterns of errors that impact rare or ambiguous words (Belinkov & Glass, 2019, p. 54). To address this issue, behavioral testing focuses on specific targeted areas of model analysis. For example, the CHECKLIST framework (Ribeiro et al., 2020) provides a method to

conduct behavioral testing focused on morphology and other areas of language to identify latent patterns of errors. Additionally, probes can also be a helpful method to identify underlying weaknesses in a model through providing visibility into what the model has successfully learned (Tenney et al., 2019, p. 4593). These principles and frameworks are applicable to assist in establishing a data-driven approach for identifying a high-risk slice (T. Wu et al., 2019, p. 747) of tokens associated with frequent and systematic fine-grained morphological tagging errors within the model. Once the slice is identified, targeted fine-tuning techniques can be applied to enhance the performance of the model.

2.8 Error Propagation in Downstream Tasks

Morphological tagging errors negatively impact downstream tasks such as parsing and translation (Tsarfaty et al., 2013, p. 20). Specifically, related to morphologically rich languages, Seeker and Kuhn observed "one of the problems when parsing morphologically rich languages…is the propagation of annotation errors in the processing pipeline and the unreliable morphological information" (Seeker & Kuhn, 2013, p. 43). Therefore, as a morphologically rich language, Ancient Greek translation into other languages also benefits from accurate morphological information (Rapacz & Smywiński-Pohl, 2025).

2.9 Architectures for Part-of-Speech and Morphological Tagging

Multiple different machine learning architectures exist to enable morphological tagging. This section provides an overview of several different models. However, an all-encompassing survey is not provided given the breadth of scope and history of the topic.

Going back decades, research was conducted related to computer-based part-of-speech tagging. Hidden Markov Models (HMMs) were proposed as a solution for

automated part-of-speech tagging (Kupiec, 1992). Although for their time, HMMs were a significant step forward in coarse-grained part-of-speech tagging ability, they came with limitations. For example, the probability of each tag depends on the tag immediately preceding it (Rabiner, 1989, p. 284). This limits the ability of the model to capture the signal through long sequences. Recurrent Neural Networks (RNNs), to include RNNs with Long-Short-Term-Memory (LSTM), provide the ability to model longer sequences as compared HMMs (Hochreiter & Schmidhuber, 1997). However, one of the disadvantages of RNNs is the signal can become undetectable over time due to the vanishing gradient problem (Bengio et al., 1994, p. 157). LSTM was developed to facilitate propagation of the signal through an arbitrary length sequence. This innovation enabled enhanced capability for part-of-speech and morphological tagging due to the ability to learn dependencies over long sequences (Plank et al., 2016, p. 412). LSTM was further enhanced, with the introduction of Bidirectional LSTM (BiLSTM) and Conditional Random Field (CRF) layers (Huang et al., 2015).

These pre-transformer architectures were largely superseded by transformer models for tagging due to the ability for transformers to be trained in parallel and the benefits of self-attention (Vaswani et al., 2017). Encoder models in particular, such as BERT (J. Devlin et al., 2019, p. 4175), were, and continue to be, widely used for tagging. Transformer language models typically begin with a pretraining stage which allows the model to learn fundamental patterns of language. Then, the model is fine-tuned on a specific subset of information to enable it to perform a specific discrete task such as tagging. Fine-tuned models such as Ancient-Greek-BERT have been shown to outperform baseline language models on fine-grained morphological tagging accuracy in

certain cases (Singh et al., 2021, p. 135). Table 3 summarizes the *overall*[6] fine-grained morphological prediction accuracy of several transformer models which are mainly pre-trained on an Ancient Greek language corpus (e.g., UD PROIEL). Ancient-Greek-BERT was selected as the primary model because it is widely adopted (Tzanoulinou et al., 2025, pp. 16–17) and demonstrates morphological tagging accuracy results (92.9% UFeats) which are slightly lower but comparable to the state-of-the-art (94.05% for greCy).

Table 3 Ancient Greek model accuracy metrics summary adapted from (Tzanoulinou et al., 2025, p. 8) and (Kostkan et al., 2023, p. 131)

Model	Dataset	Morphology (UFeats) accuracy, %	Notes
Ancient-Greek-BERT	UD PROIEL	92.90	Subword BERT pre-trained on Ancient/Medieval/Modern Greek; fine-tuned for morphological tagging
OdyCy	UD PROIEL	92.84	Pipeline built on Ancient-Greek-BERT
greCy	UD PROIEL	**94.05**	Uses a spaCy pipeline
Stanza	UD PROIEL	92.20	Stanford neural pipeline baseline
UDPipe	UD PROIEL	88.62	
GRεTA		Not reported	"Comparable" to Perseus 93.13%, but no metrics provided for PROIEL

2.10 Fine-Tuning Approaches

Fine tuning is a transfer-learning activity which begins with a pre-trained model and exposes the model to a task-specific dataset, often with additional strategies, to achieve improved results at a specific task (Howard & Ruder, 2018). As relates to transformer models, there are several approaches to fine tuning (of which only the most

[6] Overall accuracy is not the same metric as per-token accuracy.

relevant will be discussed). Generally, the approaches fall into three categories: loss, data, and architecture.

2.10.1 Loss-based Approaches

Loss-based approaches utilize a modified loss function to ensure emphasis on specific and desired areas of a task. The loss function guides the learning process within the model by determining the gap between predicted outputs and gold-standard labels. Certain loss functions adjust the learning process in how the weights are updated to yield improved results depending on the use-case.

Standard (or baseline) fine-tuning typically utilizes a cross entropy (CE) loss function (J. Devlin et al., 2019, p. 4174). CE loss is typically selected for baseline fine-tuning because it penalizes incorrect predictions where the model was confident to a greater degree than uncertain predictions, and especially categorical prediction (Connor et al., 2024). As shown in Equation 3, CE loss calculates the distance between the predicted probability and the true distribution. Standard CE-based fine-tuning has been shown to be effective in practice for many tasks, but it also has drawbacks such as under-emphasis in rare classes due to a focus on training the majority classes (Johnson & Khoshgoftaar, 2019). This under-emphasis is evident in the results of standard fine-tuning against morphologically rich languages which contain ambiguous words (e.g, Ancient Greek) (Singh et al., 2021).

$$\mathrm{CED}(q{:}p) = -\sum_{i=1}^{n} q_i \log p_i$$

Where q is the true distribution,

and p is the predicted probability distribution.

Equation 3 Cross Entropy Loss Function
(Connor et al., 2024)

Focal loss, as the name suggests, modifies CE loss and focuses the fine-tuning on the most difficult cases of class imbalance. "The loss function is a dynamically scaled cross entropy loss, where the scaling factor decays to zero as confidence in the correct class increases" (Lin et al., 2017, p. 2981). Equation 4 provides the definition for focal loss. The γ term is the hyperparameter that is tuned to control the degree of rate of down weighting. If γ is increased too high, it may overly focus on label noise instead of focusing on intermediate confidence terms (Adhikari et al., 2021). In the context of morphologically rich languages ambiguous words frequently exist in rare forms. CE loss does not typically focus on the rare cases and treats all cases as equal. Focal loss has been used to ensure that the rare cases are emphasized for additional learning in settings other than Ancient Greek (Abudouwaili et al., 2022). This suggests a potential for improved fine-grained morphological tagging within Ancient Greek.

$$\mathrm{FL}(p_t) = \ -(1 - p_t)^{\{\gamma\}} \log(p_t)$$

Where p_t is the models estimated probability (ground truth),

and γ (gamma) is the focusing parameter.

Equation 4 Focal Loss Function
(T.-Y. Lin et al., 2017)

The issue of class imbalance is widely researched across multiple domains to include computer vision and speech recognition (Johnson & Khoshgoftaar, 2019). Class-balanced loss is another approach which attempts to improve the accuracy of under-represented classes. The loss function utilizes an algorithm to re-weight via "the effective

number of samples", resulting in a loss function which is sensitive to balancing the classes (Cui et al., 2019). Optionally, additional approaches (e.g., Focal loss) can be combined with the class-balanced loss function as described in Equation 5.

$$CB_{FL}(z, y) = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} (1 - p_i^t)^{\gamma} \log(p_i^t)$$

Where n_y is the number of training samples in class y,

and β is the hyperparameter that controls the effective number of samples,

and p_i^t is the models' estimated probability of class i,

and γ (gamma) is the focusing parameter.

Equation 5 Class-Balanced Focal Loss Function
(Cui et al., 2019, p. 9264)

Although several loss functions such as focal loss and class-balanced loss are rooted in computer vision, they are generalizable to other problem spaces. Prior research has demonstrated the ability for loss functions intended for use with computer vision to be applied to the natural language processing domain. For example, the integration of focal loss with BERT-based architectures proved effective with imbalanced data (J. Liu et al., 2021).

2.10.2 Data-driven Approaches

Data-driven (or sometimes referred to as sampling) approaches select and promote the under-represented or difficult slices of data (He & Garcia, 2009). In contrast to the loss-based approaches which focus on adjusting the penalty within the loss function for inaccurate predictions, data-driven approaches attempt to increase training

on the under-represented data to increase model accuracy. Numerous data-driven approaches exist, with only the most relevant described.

Oversampling is an approach that attempts to introduce an increased number of instances of the under-represented data into the training activity. The repeated exposure to the under-represented data is intended to cause improvement to the overall learning accuracy of the specific hard cases. Synthetic Minority Over-sampling Technique (SMOTE) is a well-known implementation technique for oversampling (Chawla et al., 2002), but not the only such technique. Oversampling can produce improved accuracy, but also introduces the potential for overfitting (He & Garcia, 2009, p. 1267).

2.10.3 Architecture-based Approaches

Architecture-based approaches influence the internal structure of how the pretrained model is fine tuned. Generally, these approaches alter how the weights within the model are adjusted during training. A few examples include layer freezing (Howard & Ruder, 2018, p. 332) and contrastive learning (Gao et al., 2021). Low-Rank Adaptation of Large Language Models (LoRA) is another example of a fine-tuning method that freezes pre-trained weights and injects trainable matrices into the transformer architecture (Hu et al., 2021). However, LoRA and similar approaches focus on efficient adaptation rather than directly resolving the class imbalance or ambiguity-driven slices.

2.10.4 Iterative Error-Driven Approaches

Approaches such as Just Train Twice (JTT) (Liu et al., 2021) provide a process which trains a model, identifies errors from the trained model, and then retrains based on the error slice. The retraining utilizes an upweighting (e.g., oversampling) mechanism which exposes the training process to a higher number of instances of the error slice

compared to the non-error slice. This approach has demonstrated benefits to include reducing spurious correlations.

However, several limitations of this approach exist. JTT functions at a coarse level of reweighting, whereas Ancient Greek morphology requires a fine-grained token-level slicing approach. For example, certain errors in Ancient Greek models may be caused by data sparsity, annotation errors, or systematic errors. JTT is unable to distinguish and isolate the signal (systematic errors) from the noise as all first pass errors are treated equally in terms of upweighting. DART-FT can be viewed as a generalization of JTT. DART-FT utilizes an approach that focuses on the fine-grained feature-value pairs to identify errors, coupled with a robust process to identify a diverse high-risk slice which focuses oversampling and loss adjustments on frequent and systematic fine-grained morphological tagging errors.

Specific to the domain of Ancient Greek morphological tagging, recent work has shown that tagging accuracy can be improved through the use of morphological lexica and multi-task learning (Keersmaekers & Mercelis, 2024). However, these approaches typically optimize for overall tagging accuracy rather than directly targeting the high-risk slice. This distinction highlights the specific gap that DART-FT seeks to address.

2.11 Summary and Conclusion

This chapter reviewed the available literature for topics related to the thesis statement: *A targeted fine-tuned Ancient-Greek-BERT model will reduce mislabeling in fine-grained Ancient Greek morphological tagging.* Topics included the relevant structure and morphological background of the Ancient Greek language. Foundational relevant data sources (e.g., PROIEL) and machine learning concepts were also explored to include

fine-tuning approaches. The literature indicates that transformer models, such as Ancient-Greek-BERT, have achieved strong performance in overall baseline fine-grained morphological tagging. However, prior research also shows accurate tagging for infrequent and ambiguous words continues to remain an issue. The inaccurate tags negatively impact downstream NLP tasks which are sensitive to even small tagging errors. Further, although standard fine-tuning techniques have been applied to Ancient Greek transformer models, they focus on overall morphological tagging accuracy in aggregate.

The gap identified in the reviewed literature is the absence of research exploring targeted fine-tuning approaches applied to Ancient-Greek-BERT. There is also a gap in analysis, evaluating the impact of targeted fine-tuning on an empirically identified high-risk slice of tokens associated with frequent and systematic fine-grained morphological tagging errors within the model. This research seeks to address these gaps through identifying an empirically defined high-risk slice, applying targeted fine-tuning techniques, and evaluating the per-slice metrics.

Chapter 3—Methodology

3.1 Introduction

The gaps identified in the current literature, namely the absence of research exploring targeted fine-tuning approaches applied to Ancient-Greek-BERT and analysis evaluating the impact of targeted fine-tuning on the high-risk slice, are addressed through the methodology in this chapter. Novel contributions are provided through:

1. A data-driven approach to identify an empirically defined high-risk slice of tokens associated with frequent and systematic fine-grained morphological tagging errors.
2. A targeted fine-tuning method to improve tagging accuracy on the high-risk slice.
3. A statistical analysis demonstrating the model performance.

The methodology is intended to test support for the thesis: *A targeted fine-tuned Ancient-Greek-BERT model will reduce mislabeling in fine-grained Ancient Greek morphological tagging.*

Further, this methodology explores the degree of sensitivity of targeted fine-tuning performance to the degree of oversampling of the high-risk token slice. The following sections describe the end-to-end pipeline, data preparation and fine-tuning phases, evaluation methods and statistical testing approaches, and specific implementation details. This method establishes a foundation for the results presented in Chapter 4.

3.2 Experimental Pipeline Overview

The experimental pipeline consists of multiple controlled steps (as depicted in Figure 3) which enable a rigorous and comparative evaluation of the targeted fine-tuning results. The process begins with the normalization of the source Ancient Greek PROIEL data, utilizing three pre-defined data splits (training, validation, and test). Next, the definition of the high-risk slice is determined, then mined from the data and frozen. Then oversampling of the data occurs in the case of a targeted fine-tuning run (oversampling is omitted for the standard baseline fine-tuning run). Loss function configuration is determined based on several potential loss functions (e.g., cross-entropy, class balance, focal). The Ancient-Greek-BERT baseline model is then fine-tuned with the defined parameters. Finally, evaluation of the impact of fine-tuning is conducted against the high-risk slice and the overall test set. Further detail is provided in the following sections.

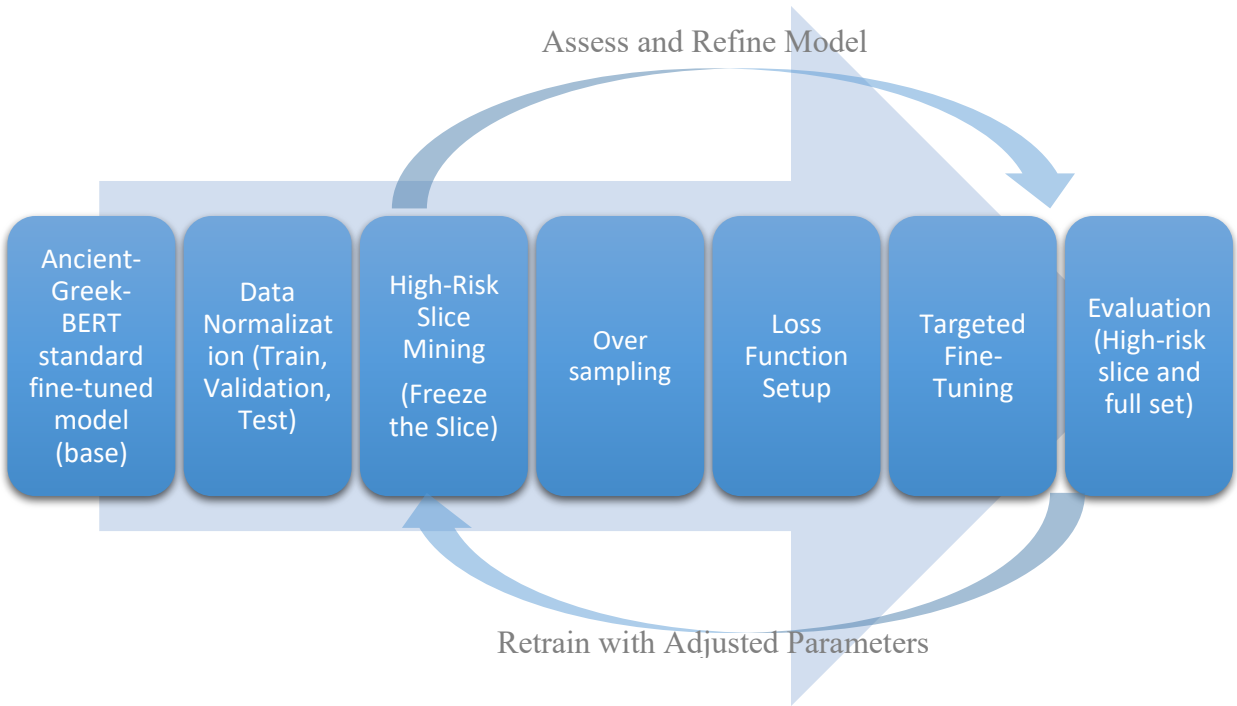

Figure 3 Pipeline depicting the end-to-end targeted fine-tuning process and evaluation

3.3 Data and Data Preparation

The Universal Dependency (UD) Ancient Greek PROIEL treebank version 2.12 is the primary data source for training, validation, and testing. PROIEL utilizes Conference on Natural Language Learning Universal (CoNLL-U) formatted files. The data is pre-split into training, validation, and testing data sets, which supports methods to prevent potential leakage across splits. All UPOS tags and UFEATS attributes are represented across all three data splits as described in Table 4.

Table 4 UD Ancient Greek PROIEL dataset composition statistics

Data split	Sentences	Tokens	UPOS tags used (count)	UFEATS attributes used (count)
Train	15016	187039	13	15
Dev (Validation)	1019	13652	13	15
Test	1047	13314	13	15

3.3.1 Data Normalization

Several data normalization steps are employed to ensure robust and standardized data is available for training and evaluation. The FEATS field contains multiple concatenations of elements delineated by a pipe (|). For example, the word λόγος has a FEATS entry of “Case=Nom|Gender=Masc|Number=Sing”. Before training, alphabetic sorting of the FEATS elements is necessary to ensure a deterministic ordering of the elements. UPOS (e.g., NOUN) is also utilized in addition to FEATS elements. Any FEATS that are empty (i.e. “_”) are replaced with an empty string (“”). If missing attributes are detected, then they are augmented by insertion of the word “NONE” for engineering purposes to explicitly note the absence of a feature.

3.3.2 Tokenization

Tokenization is achieved through typical BERT tokenization processes using WordPiece (Y. Wu et al., 2016). Each word in the corpus is decomposed into one or more subword elements. The first subword within the word is tagged with the appropriate morphological label (i.e., UPOS and FEATS). All subwords except the first subword are excluded from influencing the loss during training to prevent duplicate tagging of the same word. This promotes learning at the word level instead of at the subword level.

3.3.3 Morphological Feature Representation

The representation of the FEATS bundle is a significant design decision. In their multilingual work, Kondratyuk and Straka utilize a single string per word which includes the full bundle of FEATS (e.g., Tag1: Case=Nom|Gender=Masc). They do so because "only a small subset of the full Cartesian product of morphological features is valid, eliminating invalid combinations" (Kondratyuk & Straka, 2019, p. 2781). This may be appropriate for the multilingual case but exacerbate data sparsity in monolingual models utilizing morphologically rich languages such as Ancient Greek.

In fact, the results provided by Kondratyuk and Straka suggest a FEATS accuracy reduction when evaluating on Ancient Greek PROIEL (92.44 for UDPipe and 82.29 for UDify) which are lower accuracy scores compared to many other low morphology languages (e.g., Japanese GSD 99.98 UDPipe and 99.97 Udify) (Kondratyuk & Straka, 2019, pp. 2792, Table 7). Therefore, this praxis decomposes each FEATS bundle into the individual members for independent per feature classification (e.g., Tag1: Case=Nom; Tag2: Gender=Masc), reducing issues with data sparsity.

3.4 Identification of the High-Risk Slice

Multiple potential methods exist for determining the high-risk slice, including a predefined list of tokens, a list based on observed errors from the validation set, or a dynamic list derived during training. The high-risk slice must move beyond aggregate metrics (Ribeiro et al., 2020), be empirically defined, consistent, reproducible, and focused on the morphological labels (rather than specifically on tokens).

Figure 4 provides an overview of the high-risk slice identification methodology. A three-pass process enables the high-risk slice to be empirically defined, constrained to avoid domination by a small number of broad feature-value pairs, and then frozen to enable consistent model evaluation. The baseline threshold and diversity limit parameters were selected as conservative values with the intent of isolating systematic error patterns while maintaining a slice size which is sufficient for stable fine-tuning and evaluation. The initial step involves the baseline standard fine-tuned Ancient-Greek-BERT model predicting the correct feature-value pair tags for each token in the validation data set. Next, the feature-value pairs are compared to the gold labels to identify the candidates for inclusion in the high-risk slice. A process is then employed to ensure the diversity and fairness in the high-risk slice. The high-risk slice is then frozen for the remainder of the experiments. The validation data set is used throughout the high-risk slice identification process to avoid data leakage.

The test data set is stamped with a mask to determine the member tokens for inclusion in the high-risk slice. This use of the test data set is not training-time data leakage, as no training occurs on the test data set. The mask is applied solely as a held-out evaluation mechanism, to ensure a consistent high-risk slice to enable comparison between models. Although gold labels are leveraged to construct the mask, many

deployment scenarios may not have gold labels available. Future work could investigate a proxy method for the gold labels, such as using the predictions of a first pass model as candidates for the high-risk tokens. Finally, the oversampling occurs on the training set.

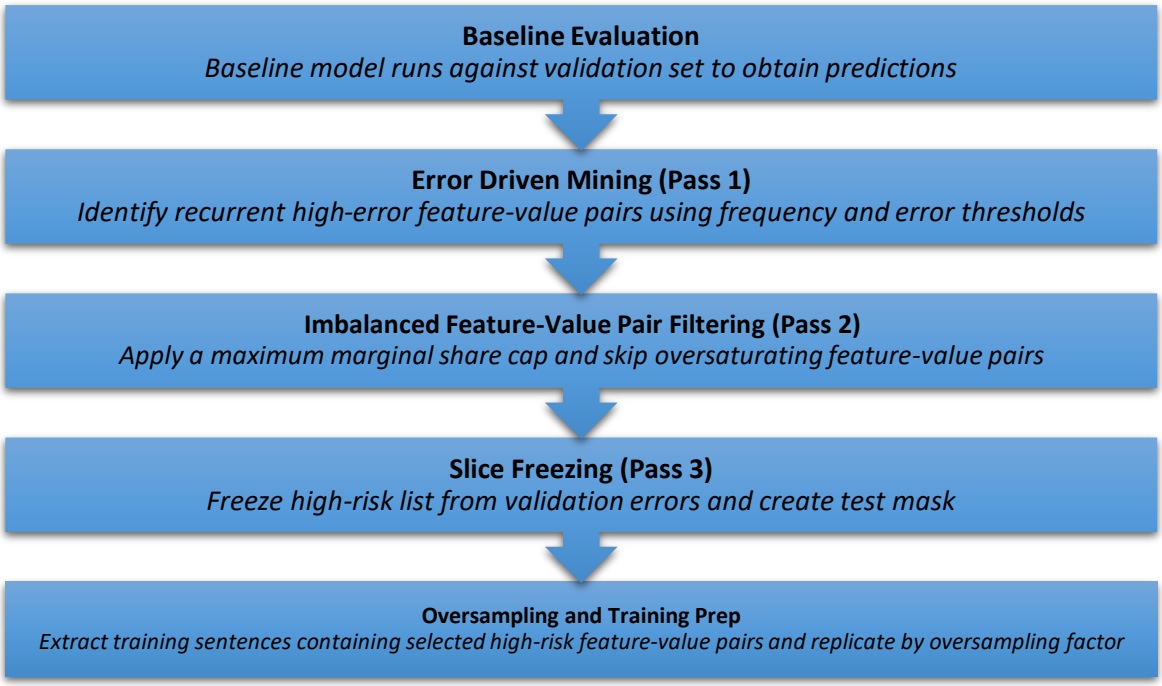

Figure 4 Overview of the methodology to identify the high-risk slice

3.4.1 Error Driven Mining (Pass 1)

The high-risk slice candidates (C_{val}), as defined in Equation 6 and partially illustrated through an example in Table 5, are determined based on the baseline standard fine-tuned Ancient-Greek-BERT model misclassification frequency ($\text{count}_{val}(f, v)$) and error rate ($\text{err}(f, v)$), both compared with a threshold (τ), on the validation set. Error patterns for tokens are determined based on the incorrect morphological label feature-value pairs (f, v) as compared to the gold labels.

$$C_{\text{val}} = \{(f, v) \in FV_{\text{val}} \mid \text{err}(f, v) > \tau_e \wedge \text{count}_{\text{val}}(f, v) > \tau_f\}$$

Where τ_e is the error rate threshold,

$\mathrm{err}(f, v)$ is the validation set error rate for the feature-value pair (f, v),

and τ_f is the minimum frequency threshold.

Equation 6 Formal definition of the high-risk slice candidates (pass 1)

Table 5 High-risk slice candidate identification example (without consideration to thresholds).

Token	Feature	Predicted Label	Gold Label	Feature-Value Error	Candidate for Inclusion in High-risk Slice?
Αὐτοῖς	Case	Dative	Dative	No	No
	Gender	Masculine	Masculine	No	No
	Number	Plural	Plural	No	No
ὁρίων	Case	Genitive	Genitive	No	No
	Gender	**Masculine**	**Neuter**	**Yes**	**Yes**
	Number	Plural	Plural	No	No

Note. Αὐτοῖς ('to them', Mark 5:19) and ὁρίων ('of the region', Mark 5:17) are both contained in the PROIEL validation set (*UD_Ancient_Greek-PROIEL Dev Split*, 2025)

3.4.2 Imbalanced Feature-Value Pair Filtering (Pass 2)

The high-risk slice can become oversaturated with certain feature-value pairs. For example, if Gender=Masculine is identified as a candidate high-risk feature-value pair, it may appear in many sentences within the training dataset. Each of these sentences would then be oversampled. Oversampling a large portion (e.g., 50-70%) of the sentences in the training dataset would reduce the potential for accuracy improvements as the training data is no longer emphasizing certain sentences over others. Additionally, training time would increase unnecessarily. Therefore, a control is needed to limit the composition of the high-risk slice to avoid oversaturation.

Two controls are implemented to prevent oversaturation:

1. Feature-value pair maximum marginal share limit: This is a primary control which limits the marginal token share that a single feature-value pair can contribute to the high-risk slice. This control is related, from a conceptual perspective to diversity and fairness mechanisms explored in other slice selection research. For example, Celis et. al. implement a fairness focused subset selection which prevents a single category of data from oversaturating the set (Celis et al., 2018), although the control differs in the specifics of implementation. Further, work to ensure diversity in a slice utilizing greedy algorithms is an active area of research (Moumoulidou et al., 2021).

   The control utilizes the validation data set to determine the set of tokens for each candidate feature-value pair. Then the control greedily builds the slice through adding one feature-value pair candidate at a time. As the high-risk slice build continues, a candidate is skipped if its marginal share exceeds the configured threshold of the high-risk slice composition. This allows for the candidate feature-value pair to be represented in the high-risk slice without oversaturating it.

   a. An illustrative example provides further clarity on the mechanics of the control.

      i. Assume that a feature-value pair maximum marginal share limit is configured to 35%. The control sorts all the feature-value pair candidates from the validation set based on their error counts.

      ii. Assume the most error-prone feature-value pair candidate is Voice=Passive. This feature-value pair would be accepted in the

slice. Also, assume it has 500 tokens within the validation data set. This would imply that the high-risk slice is now 500 tokens in size.

iii. Assume the next most error-prone feature-value pair candidate is Case=Nominative. It also has a large footprint but only contains 50 tokens which are not accounted for in the previous step with the feature-value pair Voice=Passive. Therefore, the high-risk slice grows to a size of 550 tokens. Before finalizing the addition of the 50 tokens from the Case=Nominative feature-value pair, the marginal share is calculated to determine if the marginal share limit (i.e., 35%) is exceeded. In this case the 50 tokens represent about 9% of the slice (50/550) resulting in the Case=Nominative tokens added to the high-risk slice.

iv. Assume the next most error-prone feature-value pair candidate is Gender=Masculine, which could contribute 375 new tokens to the high-risk slice. This contribution would oversaturate the high-risk slice and cause data imbalance with a contribution of approximately 40% (375/925). Therefore, the Gender=Masculine feature-value pair is not included in the high-risk slice. The slice size remains at 550 tokens.

v. This process continues until all the feature-value pair candidates are either included in the high-risk slice or omitted.

2. Feature-value pair exclusion: This is a secondary control which removes a specified feature-value pair candidate from consideration for oversampling.

Both controls are optional depending on the parameters of the feature-value pair candidate selection. In this specific set of research experiments, the feature-value pair maximum limit is utilized at 35%, and the feature-value pair exclusion is not utilized.

3.4.3 Slice Freezing and Mask Creation (Pass 3)

The input for slice freezing is the output from passes 1 and 2. Pass 1 provided the candidate feature-value pairs and defined in C_{val}. Pass 2 defined a subset of the candidates $S^{\star} \subset C_{\text{val}}$ necessary to ensure fairness and diversity in the high-risk slice. Pass 3 applies a mask $M(S^{\star})$, as shown in Equation 7, to the test data set to define the high-risk slice. This frozen mask is used only for held-out evaluation and consistent model comparison.

$$S_{risk} = M(D_{test}, S^{\star})$$

Where M is the masking function,

and D_{test} is the test dataset,

and $S^{\star}$ is the subset of the candidate feature-value pairs.

Equation 7 High-risk slice is frozen based on the filtering candidates (pass 3)

3.5 Fine-tuning Methodology

The targeted fine-tuning methodology utilizes a base model (i.e., Ancient-Greek-BERT standard fine-tuned) and further enhances the base model to increase accuracy on the high-risk slice. Several parameters are modified with each experiment, and several other parameters remain constant to support reproducibility. The primary constant parameters are the learning rate, batch size, and number of epochs, as described in Table 6. The primary variable parameters, as described in Table 7, are the oversampling factor and the loss functions. Multiple seeds are applied for each experimental run to

further provide a robust experimentation environment and limit the chance of random behavior implying statistically significant results.

Table 6 Constant Parameters used during the training

Constant Name	Value
Learning Rate	$3x10^{-5}$
Batch Size	8
Epochs	5
Optimizer	AdamW
Weight Decay	0.01
Maximum Sequence Length	512
Mixed Precision	Enabled (fp16)
Dropout	0.1
Learning Rate Scheduler	None

Table 7 Variable Parameters used during the training and evaluation

Variable Type	Name	Description	Values Evaluated
Independent	β (Beta)	Class balanced smoothing parameter	0.991, 0.992, 0.993
Independent	γ (Gamma)	Focal loss focusing parameter	1.5, 2.0
Independent	OS	Oversampling Factor	2x, 3x, 4x
Dependent	Accuracy on High-risk Slice	Accuracy on high-risk token slice	0-100%
Dependent	Global F1	Global macro F1 score	0-1.0

3.5.1 Baseline Standard Fine-Tuning

Establishing the standard fine-tuned model to compare the targeted fine-tuned models against is important to maintain a consistent and robust experimental pipeline.

The standard fine-tuned model utilizes Ancient-Greek-BERT and applies a cross-entropy loss function (as defined in Equation 3) against the training set during the fine-tuning process. Oversampling on the high-risk slice is not performed during standard fine-tuning.

3.5.2 Targeted Fine-Tuning

The targeted fine-tuning methodology employs a combination of several experimental variables related to oversampling and loss functions. Random oversampling with replacement (Japkowicz & Stephen, 2002) is an approach that increases the frequency of certain training data and can be a straightforward and effective means of targeted fine-tuning. The oversampling factor determines the rate at which the high-risk slice is oversampled throughout the training set to enhance learning visibility. The oversampling factor of the high-risk slice as an independent variable with a value of 2x, 3x, or 4x. The targeted fine-tuning process utilizes the baseline standard fine-tuned model (with the corresponding random seed) for a warm initialization for transfer learning.

3.5.2.1 Cross Entropy Loss with Oversampling

The loss function type is also variable, including the application of cross-entropy, class-balanced, and class-balanced combined with focal loss. The cross-entropy loss function, as defined in Equation 3 is the same function used with the standard fine-tuning baseline models with the addition of an oversampling factor of 2x, 3x, or 4x.

3.5.2.2 Class Balanced Loss with Oversampling

The class-balanced loss function uses the Effective Number of Samples method (Cui et al., 2019) as defined in **Equation 8**. The method seeks to ensure the emphasis on

majority classes is reduced in long-tail data sets by accounting for redundancy of samples within the same class. This approach is well suited to the Ancient Greek morphological ambiguity enhancement problem space given the large number of repeated tokens within the high-risk slice. In this implementation, a class is a specific morphological feature-value pair.

$$w_c = \frac{1-\beta}{1-\beta^{n_c}}$$

Where

w_c is the class-balanced weighting,

β is a smoothing hyperparameter which controls redundancy in a class,

n_c is the number of samples for the class.

Equation 8 Class Balanced Weight function
(Cui et al., 2019, p. 9263)

β is an independent variable which is represented as $\frac{(N-1)}{N}$ where "N can be interpreted as the number of unique prototypes" (Cui et al., 2019, p. 9266). In coarse-grained datasets, such as with Ancient Greek UPOS tagging, N (which represents internal diversity within a class) is usually large. This results in a β value approaching 1. However, the Ancient Greek morphological data is fine-grained which suggests N should be a modest value (e.g., 100-150). Therefore, the values of β were selected in a targeted range of 0.991 - 0.993 given the highly sensitive relationship between N and β. For example, a value of β =0.991 corresponds to an N of approximately 112, and β = 0.993 corresponds to N of approximately 143.

3.5.2.3 Class Balanced Loss with Focal Loss and Oversampling

When the focal loss function is utilized, as defined in Equation 4 , the parameter (γ) is an independent variable which reduces the weights for the easier examples. γ is iterated with the values of 1.5 and 2.0 during the experimental runs. The values of 1.5 and 2.0 were selected due to the strong performance of γ =2 observed by Lin et al. (T.-Y. Lin et al., 2017, p. 3004) and an alternative value of 1.5 to test the sensitivity with a slightly less aggressive focal function. When both class-balanced and focal loss are combined, w_c (as defined in **Equation 8**), is multiplied by the focal loss term.

3.6 Evaluation Metrics and Statistical Testing

Once the training step in the fine-tuning process is complete, prediction and evaluation steps occur. The prediction process utilizes the newly fine-tuned model and the test data set to generate the predicted morphological labels. The predicted labels are aligned with and compared to the gold labels from the PROIEL dataset, which allows for the computation of the overall and high-risk slice accuracy and F1 scores. A Boolean mask is utilized to flag which tokens are included in the high-risk slice. This mask ensures a consistent held-out definition of the high-risk slice across models for comparison purposes.

3.6.1 Evaluation Metrics

The primary evaluation metrics include bundle level accuracy and macro F1 score. The bundle level accuracy calculation is determined by the exact match of the feature-value pairs within the UPOS and full set of morphological features within the FEATS bundle. If there is an exact match between the full bundle and the gold label, then

100% accuracy is achieved for that token. If there is a partial match or no match between the full bundle and the gold label, then 0% accuracy is achieved for that token.

Jeni et al. provide a definition for F1 score. The F1 score, as expressed in Equation 9, is the harmonic mean of the precision and recall values (Jeni et al., 2013, p. 248).

$$\mathrm{F}_1 = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

Where

P is precision,

R is recall.

Equation 9 Formal definition of F1 score metric (Jeni et al., 2013, p. 248)

As defined in Equation 10, a macro F1 score is utilized. The macro F1 score averages each class equally regardless of the frequency. Some morphological feature-value pairs can be sparse in the test set. Therefore, it is essential to use macro averaging to prevent frequent feature-value pairs from dominating the F1 score.

$$maF_1 = \frac{1}{N}\sum_{i=1}^{N} F_{1,i}$$

Where

N is the number of classes (unique feature-value pairs),

F_1 is the F1 score for class i

Equation 10 Macro F1 score definition (Takahashi et al., 2022a, p. 4963)

Although accuracy is intuitively well understood, there are reasons for caution for heavy reliance in certain circumstances. Specifically, accuracy can be a misleading metric when strong class imbalance exists (Saito & Rehmsmeier, 2015, p. 7/21, Table 2). Ancient Greek contains imbalance due to the morphological richness of the language, where certain combinations of morphological features exist far more frequently than others.

Evaluation of the targeted fine-tuned model is viewed through two lenses:

1. Performance on the high-risk slice
2. Overall performance across the full set of data

These two views on the performance of the model allow for a deeper understanding of the impacts of the targeted fine-tuning processes. The metrics describing the performance of the high-risk slice (as defined in Equation 7), demonstrate the specific impacts of the targeted fine-tuning process on the high-risk slice which is essential to supporting the thesis: *A targeted fine-tuned Ancient-Greek-BERT model will reduce mislabeling in fine-grained Ancient Greek morphological tagging*. The overall performance metrics are also essential to support understanding of Hypothesis 1: *Targeted fine-tuning does not negatively impact global macro F1 performance compared with a standard fine-tuned model.*

3.6.2 Statistical Testing

Statistical testing provides support to demonstrate that the changes between the baseline standard fine-tuned and targeted fine-tuned models are not by random chance. Instead, they occur as a result of the targeted fine-tuning process and are therefore statistically significant. Multiple statistical testing options exist depending on the data

distribution. Some illustrative examples of common statistical testing methods include t-tests (Student, 1908) and McNemar (McNemar, 1947). Although t-tests are commonly used, they have been shown to be suboptimal for the paired binary type of testing because they are designed for continuous data and can have high false-positive error rates when utilized with noncontinuous data (Dietterich, 1998). The two-sided exact McNemar test and Confidence Intervals are also utilized.

3.6.2.1 McNemar test

The McNemar test works through utilizing a 2x2 matrix as illustrated in Table 8, which compares the predictions of the baseline and targeted fine-tuned models against the same set of tokens. The McNemar test highlights occurrences of disagreement between the models. n_{10} represents the number of occurrences with the baseline prediction incorrect and the targeted fine-tuned model correct, and n_{01} represents the number of occurrences with the baseline prediction correct and the targeted fine-tuned model incorrect.

Table 8 McNemar 2x2 matrix

Model Outcome	Targeted Fine-Tune Predicted Correct	Targeted Fine-Tune Predicted Wrong
Baseline Correct	—	n_{01}
Baseline Wrong	n_{10}	—

The exact McNemar test is utilized instead of the originally proposed Chi-squared distribution (McNemar, 1947) due to the size of the data. The Chi-squared distribution has been shown to be unreliable and inaccurate with small datasets (Liddell, 1983). The high-risk slice, depending on the specific pipeline run, may fall into the category of a small dataset. Additionally, Fagerland indicates that the exact McNemar test can be too

conservative in some cases and strictly controls the false-positive (i.e. Type 1) error rate (Fagerland et al., 2013). However, a robust and reliable test which might be considered too conservative in some cases is desired to reduce the chances of false positives. A two-sided test is used as the possibility for an increase or a decrease in correct predictions exists. Therefore, to support a robust and reliable test, the exact two-sided McNemar test is utilized as depicted in Equation 11. The $\min(n_{12}, n_{21})$ in the equation is the smaller (minimum) of either n_{01}or n_{10}. n is the number of discordant pairs. In terms of implementation, the scipy.stats library (Virtanen et al., 2020) with the exact two-sided binomial test is used.

$$p = 2 \sum_{x_{12}=0}^{\min(n_{12},n_{21})} \binom{n}{x_{12}} \left(\frac{1}{2}\right)^{n}$$

Where n_{12} is the baseline correct / targeted fine-tuned incorrect,

n_{21} is the baseline incorrect / targeted fine-tuned correct,

n is the total number of discordant pairs.

Equation 11 Exact Two-sided McNemar Test (Fagerland et al., 2013, p. 3)

3.6.2.2 Confidence Intervals

Confidence Intervals (CIs) provide a method to validate the reliability of the accuracy change. The Wald interval is a commonly used approach in CIs. However, Agresti cautions against the use of the Wald intervals with small sample sizes (Agresti, 2019, p. 230) such as with the high-risk slice. An alternative but related approach is the

Adjusted Wald, or sometimes referred to as the Wald+ 2 method, for matched pairs (Agresti & Min, 2005). The Wald+2 method modifies the standard error (SE) formula as shown in Equation 12, to "add 2 observations before applying the Wald formula, by adding 0.5 to every cell of the 2 x 2 table." (Agresti, 2019, p. 230)

$$SE = \frac{1}{n+2}\sqrt{(n_{12} + n_{21} + 1) - \frac{(n_{12} - n_{21})^2}{n+2}}$$

Where

n_{12} is the baseline correct / targeted fine-tuned incorrect,

n_{21} is the baseline incorrect / targeted fine-tuned correct,

n is the total sample size of the high-risk slice.

Equation 12 Agresti-Min Wald+2 Confidence Interval Standard Error (Agresti, 2019, p. 230; Agresti & Min, 2005)

The 95% CI as noted in Equation 13, utilizes the Agresti-Min Wald+2 CI Standard Error (SE) from Equation 12.

$$CI = \frac{n_{21} - n_{12}}{n+2} \pm 1.96 \times SE$$

Where

n_{12} is the baseline correct / targeted fine-tuned incorrect,

n_{21} is the baseline incorrect / targeted fine-tuned correct,

n is the total sample size of the high-risk slice.

Equation 13 Agresti-Min Wald+2 95% Confidence Interval

3.6.3 Multi-seed

Numerous researchers have identified that variations occur between runs due to non-deterministic variables. Therefore, multiple seed values are necessary to ensure reliable results in many areas of machine learning (Reimers & Gurevych, 2017) (Henderson et al., 2018). This principle applies to transformers and BERT models specifically. Multiple seeds in the pipeline are utilized.

3.7 Implementation Details

The specific hosting platform used is the Google Cloud Vertex AI. A single g2-standard-4 (Graphics Optimized: 1 NVIDIA L4 GPU, 4 vCPUs, 16GB RAM) machine, environment version M133, was utilized in the us-east4-a zone. JupyterLab 3 was utilized, in conjunction with a GitHub repository for source code configuration management.

3.8 Downstream Translation Pilot (PROIEL Ancient Greek to Latin)

A small-scale and limited scope downstream proof of concept pilot evaluation was conducted to determine if improvements in morphological tagging from DART-FT result in changes to translation quality. Overall, the pilot injected DART-FT refined Ancient Greek morphological information into the Ancient Greek source text, which the model then translated into corresponding Latin. The pilot utilized both the Ancient Greek and Latin PROIEL treebanks, with a limited corpus consisting of only the New Testament to ensure parallel synchronization. The “Ref=” identifier in PROIEL was utilized to ensure verse-by-verse alignment between the Ancient Greek and Latin (e.g., Ref=Matthew 1:1).

The translation process utilized the Helsinki-NLP/Opus-MT-TC-bible-big-mul-mul (multilingual) model pre-trained on many languages (to include but not limited to Ancient Greek and Latin New Testament text) (*Helsinki-NLP/Opus-Mt-Tc-Bible-Big-Mul-Mul · Hugging Face*, 2024). In separate experiments for a basis of comparison, morphology information from the best performing DART-FT (seed 43, 3x oversampling factor, class-balanced loss, β=0.993) baseline and models was injected (e.g., <UPOS=Propn>, <Case=Nom>) to provide morphological context to the translator without causing subword fragmentation. The injected morphological features were only selected from the train and validation sets of the Ancient Greek PROIEL treebank to avoid leakage. If a unique feature was later discovered when processing the test set, the <UD_UNK> tag was used. The metrics to evaluate translation quality were character n-gram F-score++ (chrF++) (Popović, 2017) and translation edit rate (TER) (Snover et al., 2006), with paired bootstrap resampling for significance testing. The rich morphology in Ancient Greek led to the selection of chrF++ over BLEU. Randomized significance tests are more reliable for machine translation quality than standard parametric statistical tests, thus the selection of paired bootstrap resampling (Koehn, 2004). Google Colab, with a G4 GPU instance (NVIDIA RTX PRO 6000 Blackwell), was utilized to execute the pilot. The pilot provided initial evidence of the downstream impacts of improved morphological tagging accuracy on the high-risk slice, but it was not a core contribution.

Chapter 4—Results

4.1 Introduction

This chapter provides the results from experiments conducted with the methodology defined in Chapter 3—Methodology. The goal of the experiments is to evaluate the thesis: *A targeted fine-tuned Ancient-Greek-BERT model will reduce mislabeling in fine-grained Ancient Greek morphological tagging.* To evaluate the thesis, several hypotheses were presented:

H1: *Targeted fine-tuning does not negatively impact global macro F1 performance compared with a standard fine-tuned model.*

H2: *Targeted fine-tuning yields significantly higher fine-grained morphological tagging accuracy for the high-risk slice compared to a standard fine-tuned Ancient-Greek-BERT model.*

H3: *High-risk slice performance is sensitive to the oversampling factor, with intermediate oversampling achieving the highest mean accuracy improvement among tested factors.*

The pipeline trained and evaluated a total of 90 models. Five unique random seeds were utilized to reduce the potential for results to be influenced from random noise. Each seed was used to train models using three oversampling configurations (i.e., 2x, 3x, 4x). The selected loss functions (i.e., class-balanced focal, class-balanced, and cross entropy) were utilized with each oversampling factor configuration. This resulted in 18 configurations per seed and a total of 90 trained models (5 seeds x 18 configurations per seed).

4.2 High-Risk Slice Characteristics

The high-risk slice is composed of the Ancient Greek tokens for which the baseline model exhibits frequent and systematic fine-grained morphological tagging errors. As defined in the methodology, the slice is constructed through a multi-pass filtering process which identifies the systematic errors. This focused approach allows for the targeting of the most important errors rather than targeting all errors indiscriminately.

4.2.1 Filtering Process Results (Passes 1-3)

As depicted in Figure 4, the pipeline utilizes a three-pass approach to act as a 'funnel' to distill the most systematic error patterns into the high-risk slice. The process was executed on each seed to ensure that the specific stochastic error patterns from the seed-specific baseline model were correctly identified. The details for each pass are summarized in Table 9 and include:

Table 9 Threshold parameters utilized in creation of the high-risk slice

Threshold Parameter	Value
Minimum Error Rate (τ_e)	**5%**
Minimum Frequency (τ_f)	**5**
Candidate FV Pair Maximum Cap	**2,000**
Diversity Limit	**35%**

- Pass 1 (Error-driven mining) – The first pass identified the morphological feature-value (FV) pairs which exceeded the defined thresholds. Specifically, the candidates were required to have a minimum error rate of 5% (τ_e) and a minimum occurrence count of 5 (τ_f) in the validation set. In addition to the threshold parameters, an implementation cap of 2,000 candidate feature-value pairs was applied as a practical safeguard to ensure the slice did not become

unmanageably large or introduce significant levels of noise. However, given the significantly smaller number of possible FV pairs within the PROIEL dataset, the cap was never reached. An average of 13.8 feature-value pairs were identified for inclusion across the five seeds.

- Pass 2 (Imbalanced feature-value pair filtering) – To ensure no single morphological feature dominated the high-risk slice, a maximum marginal share limit control was implemented. The control stipulated that if a morphological feature-value pair contributed more than 35% of the high-risk slice, it should not be included in the high-risk slice. This approach supported a diverse and focused high-risk slice composition. Logs indicated that this control was active, as it rejected an average of 6.4 over-dominate feature-value pairs from pass 1 (which was approximately 46% of the pass 1 candidates).
- Pass 3 (Slice freezing and masking) – The final pass applied the filtering criteria to the dataset, resulting in an average of 7.4 feature-value pairs. A Boolean mask was applied to freeze the slice and ensure the targeted fine-tuning and evaluation were applied consistently. This resulted in a high-risk slice average of 574 tokens.

The threshold values used were fixed design choices selected before the main experimental runs. Section 4.4.3 evaluates the sensitivity of the threshold choices through an ablation study. The values selected produced a stable slice for experimentation but were not exhaustively optimized across all possible potential values. The thresholds may significantly impact the identification of the high-risk slice and resulting fine-tuning outcomes for future applications. Finally, the quantitative high-risk slice identification results, across the 5 seeds, are presented in Table 10 and Table 11.

Table 10 High-Risk Slice development for each seed through each pass

Seed	Pass 1 Candidates (FV Pairs)	Pass 2 Rejected (FV Pairs)	Pass 3 Retained (FV Pairs)	Final Slice Size (Tokens)
41	12	6	6	487
42	15	6	9	639
43	14	6	8	621
44	14	8	6	487
45	14	6	8	636
Avg	**13.8**	**6.4**	**7.4**	**574**

Table 11 Summary metrics for each pass averaged across the five seeds

Filtering Stage	Metric	Average Count	Impact / Reduction
Pass 1	Candidate Error Patterns (FV Pairs)	13.8 FV Pairs	*Initial Capture*
Pass 2	Rejected (FV Pairs)	6.4 FV Pairs	46% Rejected (Diversity Filter)
Pass 3	Filtered Error Patterns (FV Pairs)	7.4 FV Pairs	Final FV Pair Selection
Final Slice	Final High-Risk Slice Size	574 Tokens	4.3% of Test Set Tokens*

*Note – The percentages are based on test set token count of 13,314 tokens.

4.2.2 Slice Composition

The composition of the high-risk slice was dynamic based on the stochastic influences from each random seed. As noted in Table 10, the range of the number of tokens within the high-risk slice spanned from 487 to 639. These variations demonstrated that the pipeline was able to adapt as different conditions were encountered.

The specific morphological features which were selected for the high-risk slice varied per seed. Table 12 provides an overview of which categories of features were selected. The values in Table 12 represent the number of feature-value pairs in each

feature category which were included in the high-risk slice. Across all five seeds, Case, Gender, Possession, and PronType were included in the high-risk slice.

Table 12 Number of feature categories included in the high-risk slice for each random seed

Feature Category	Seed 41	Seed 42	Seed 43	Seed 44	Seed 45	Average
Case	1	1	1	1	1	**1.0**
Gender	1	1	1	1	1	**1.0**
Possession	1	1	1	1	1	**1.0**
PronType	1	1	1	1	1	**1.0**
Mood	2	2	1	-	2	**1.4**
Tense	-	2	2	1	1	**1.2**
Degree	-	1	1	1	1	**0.8**
Total Features	**6**	**9**	**8**	**6**	**8**	**7.4**

The total theoretical number of feature categories for potential inclusion in the high-risk slice is 16 (15 FEATS morphological features + 1 UPOS) and feature-value pairs 58 (45 FEATS + 13 UPOS tags) which can be derived from a summation of the published Ancient-Greek PROIEL treebank morphology information (*UD Ancient Greek PROIEL*, 2025). The visualization in Figure 5 identifies the areas of feature category inclusion and in Figure 6 identifies the specific feature-value pair inclusion in the high-risk slice across the 5 random seeds.

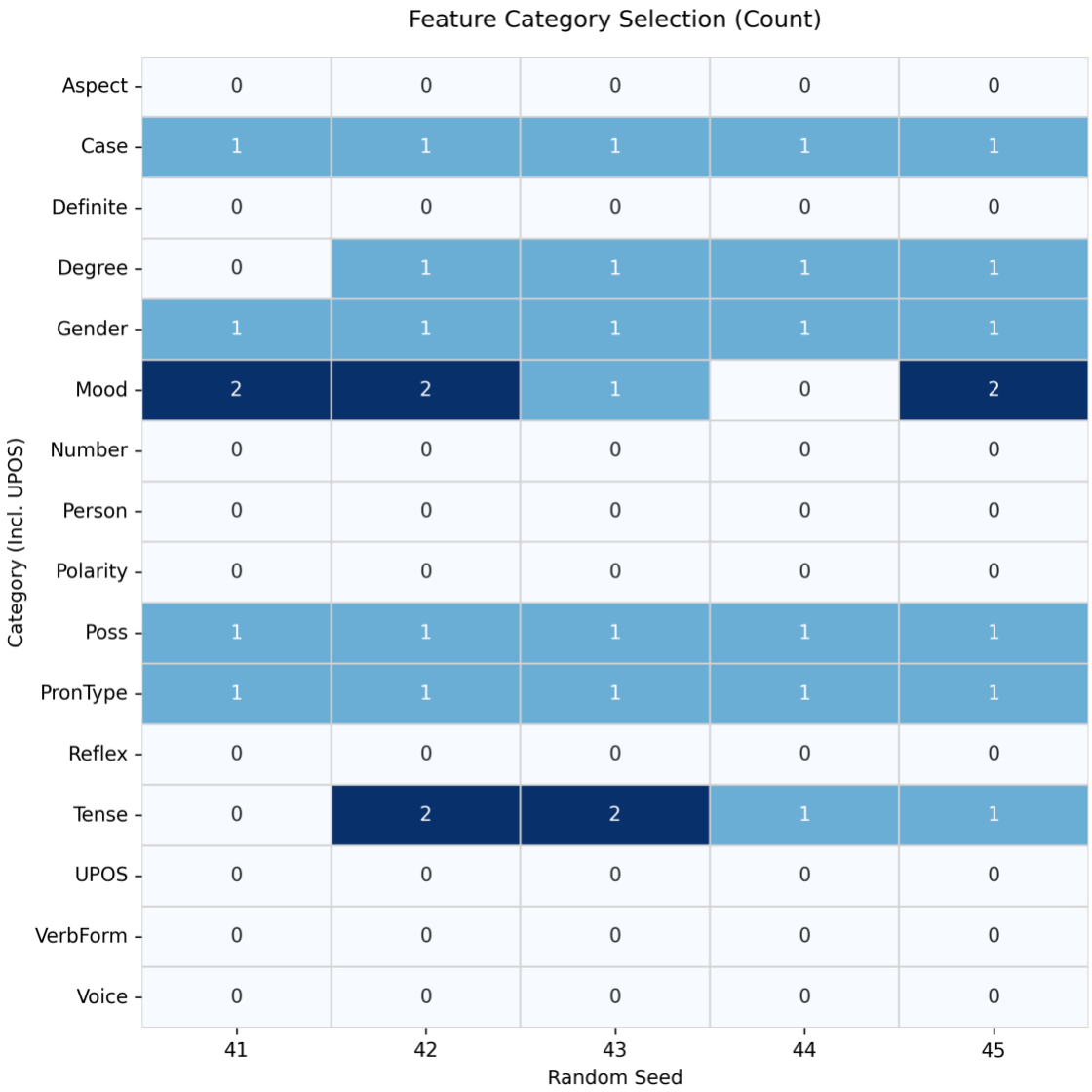

Figure 5 Feature category inclusion in the high-risk slice across the 5 random seeds

Figure 6 Feature-value pair inclusion in the high-risk slice across the 5 random seeds

Once the final list of high-risk feature-value pairs is selected, the tokens containing the selected feature-value pairs are included in the high-risk slice. Each feature-value pair contributes a variable proportion of tokens to the slice. Figure 7 visualizes the composition of the high-risk slice across the 5 random seeds. It is noteworthy that the Gender=Masc,Neut feature-value pair is dominant across all 5 seeds, ranging from 32.2% to 42.3% of the total unique tokens in the high-risk slice. Other feature-value pairs were less consistent in their dominance depending on the random seed initialization. However, Mood=Sub (29.8%, seed 41), Tense=Fut (28.3%, seed 44), and Mood=Imp (18%, seed 45) also contributed significantly to the high-risk slice under certain seeds.

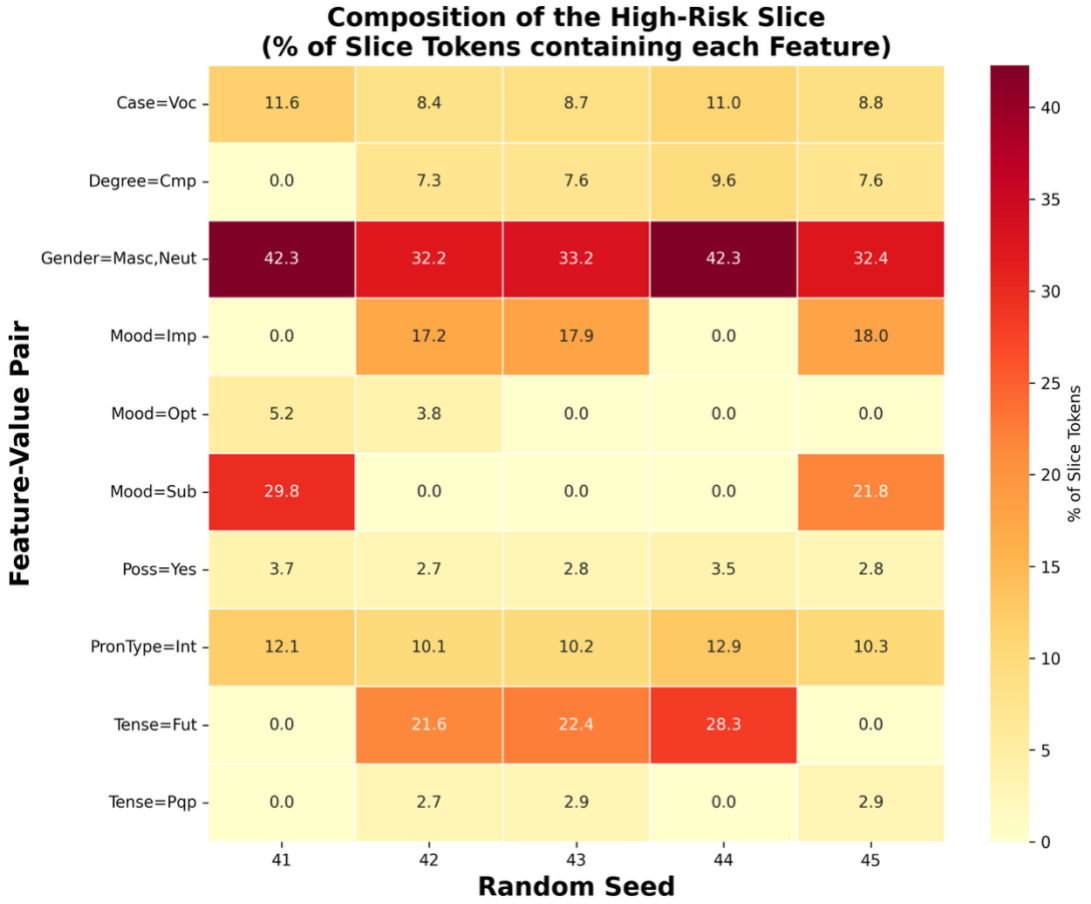

Figure 7 Feature-value pair composition of the high-risk slice (Note – column totals slightly exceed 100% because a single ambiguous token can possess multiple high-risk feature-value pairs simultaneously)

4.2.3 Token Level Composition

The feature-value pair analysis provides a high-level understanding of the sources of errors. However, analyzing the specific tokens included in the slice provides a clear linguistic connection and interpretation of the errors. Figure 8 illustrates which Ancient Greek tokens are included in the high-risk slice across all 5 random seeds. Tokens included in the high-risk slice represent errors which the baseline fine-tuned model exhibited, rather than a claim related to linguistic ambiguity.

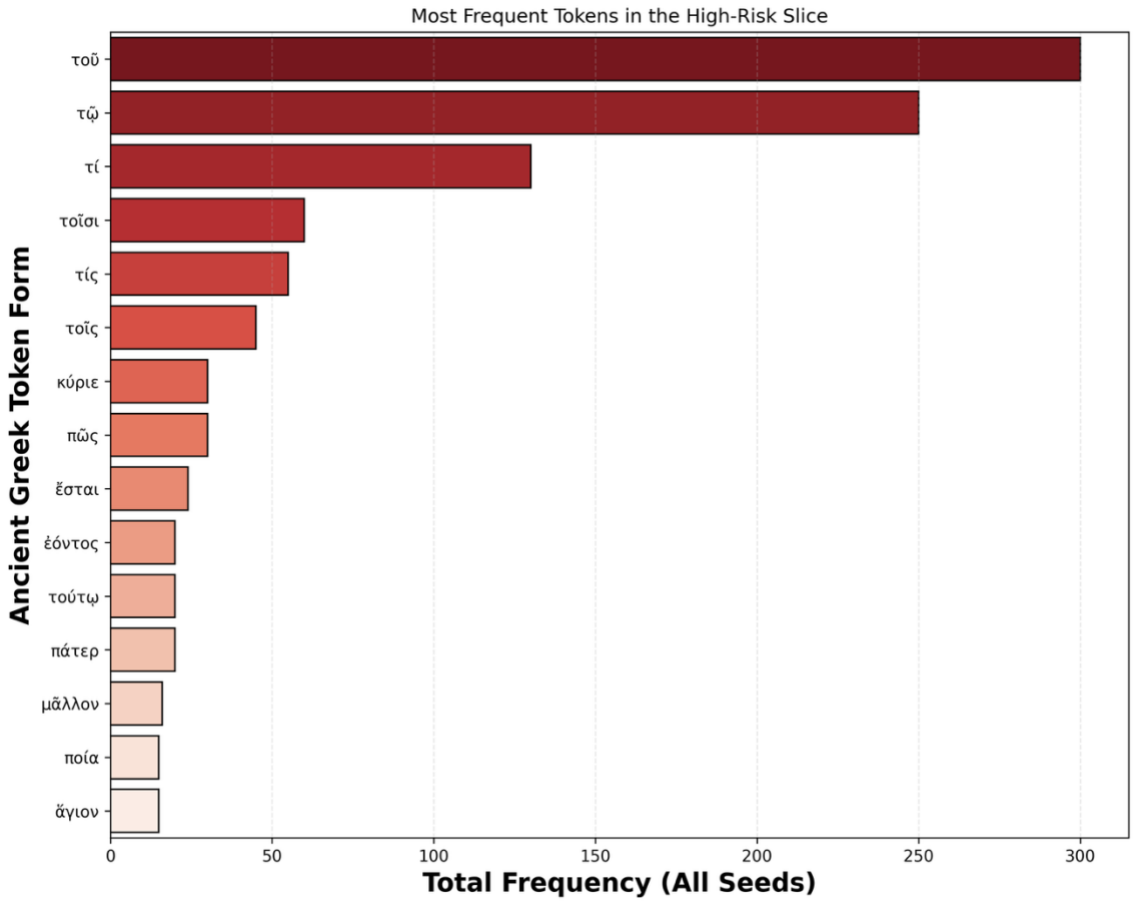

Figure 8 Frequency of the most common Ancient Greek tokens in the high-risk slice across all 5 random seeds. Tokens represent instances where the baseline model failed to accurately predict fine-grained morphology.

4.3 High-Risk Slice Performance

With research question and hypothesis 2 in mind, this section provides results related to the performance of the high-risk slice predictions of the targeted fine-tuned models.

4.3.1 Evaluation Setup

The performance of the targeted fine-tuned model was evaluated through a direct comparison of the performance of the baseline standard fine-tuned model. Models with the same random seed were paired for comparison to ensure a consistent environment for

evaluation. For example, if the baseline standard fine-tuned model was trained and evaluated with a random seed of 41, the targeted fine-tuned model under comparison was also trained and evaluated with a random seed of 41. Two primary metrics were utilized to derive quantitative information. The fine-grained token accuracy and the Macro F1 score provide insight into the performance. Accuracy is the exact match between all the fine-grained morphological features (i.e., FEATS) and UPOS as compared to the gold labels. Partial matches are considered failures. Macro F1 is the harmonic mean of the precision and recall for each feature-value pair individually and then averaged without weighting, as defined in Equation 10. The focus of these metrics is specifically against the high-risk slice rather than on other slices or globally.

4.3.2 Quantitative Performance Improvements

The targeted fine-tuning method, as described in Chapter 3—Methodology, demonstrated performance gains for the best configurations across experiments utilizing all 5 random seeds. Table 13 provides the fine-grained accuracy and Macro F1 gains. Figure 9 provides a comparison between the fine-grained morphological accuracy on the baseline model and the fine-tuned model across seeds.

Macro F1 scores showed consistent increases as compared to baseline across the configurations. Exact match accuracy gains were especially robust and significant as compared to baseline configurations, with the best configuration for seed 43 yielding a 6.4 percentage point accuracy gain on the high-risk slice. The baseline F1 scores across all seeds were relatively high as a starting point, which indicates that the baseline models were successfully predicting most feature-value pairs across the full bundle for tokens in

the high-risk slice. The sharp rise in accuracy indicates that the most difficult errors within the bundle were resolved through the targeted fine-tuning method.

Table 13 Comparison of baseline fine-tuned and targeted fine-tuned models on the high-risk slice for the highest performing models for each random seed

Seed	Baseline Accuracy	Targeted Accuracy	Accuracy Change (pp)	Significance (p)	95% Confidence Interval	Baseline Macro F1 (high-risk slice)	Targeted Macro F1 (high-risk slice)	Macro F1 Change
41	59.3%	63.9%	+4.5	< .001	1.29-7.71	0.873	0.884	+0.011
42	65.6%	70.7%	+5.2	.003	2.37-7.92	0.890	0.893	+0.003
43	63.4%	69.9%	+6.4	< .001	3.38-9.14	0.883	0.888	+0.005
44	57.5%	63.2%	+5.7	< .001	1.79-9.25	0.860	0.880	+0.020
45	64.5%	69.8%	+5.3	< .001	2.67-7.99	0.866	0.880	+0.014

*Note – Differences utilize unrounded values

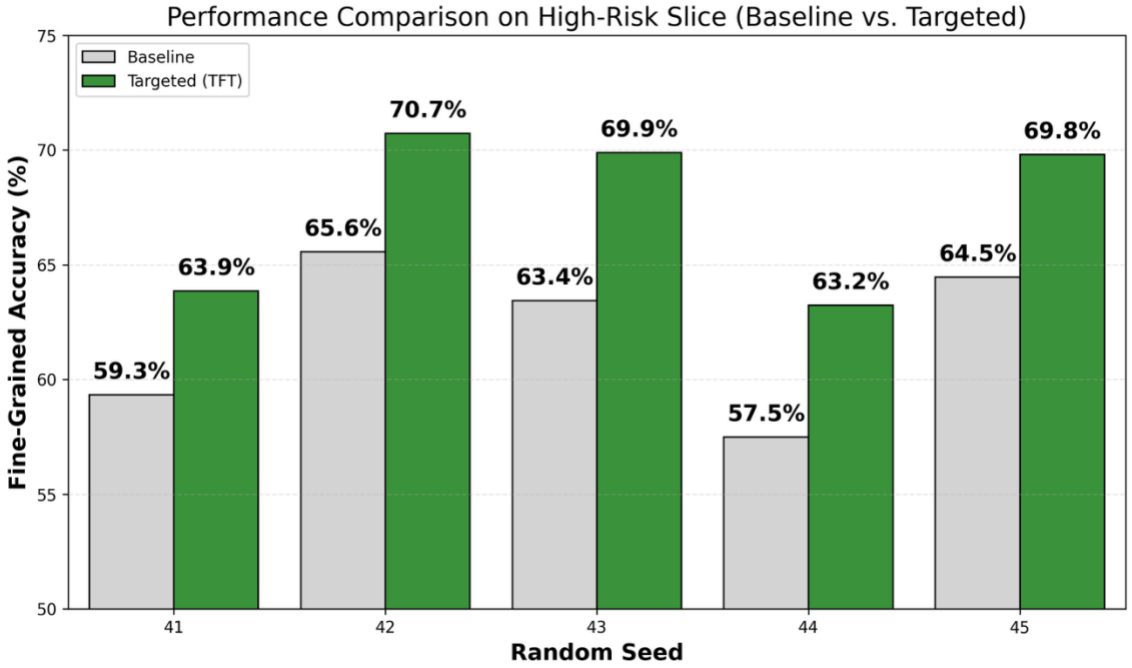

Figure 9 Fine-grained morphological accuracy comparison between baseline and fine-tuned models for the highest performing models for each random seed

4.3.3 Error Correction Analysis

An error correction analysis was performed to identify which specific tagging errors were corrected via the targeted fine-tuning method. The analysis examined the tokens where the baseline model prediction was incorrect, and the targeted fine-tuned model was correct. Table 14 provides a selection of example instances for illustrative purposes.

The analysis of the corrections shows that the errors in the morphologically ambiguous feature-value pairs were substantially reduced through the targeted fine-tuning. Figure 10 shows that 57.6% of the corrections of the highest performing models across all 5 seeds involved the Gender=Masc,Neut feature-value pair. The baseline model tended to default to a majority class (e.g., Gender=Masc or Gender=Neut) instead of appropriately predicting Gender=Masc,Neut. The resolution of the Gender=Masc,Neut errors indicates the targeted fine-tuning enabled the models to correct a recurring error pattern observed in the baseline models.

Table 14 Illustrative examples of token fine-grained morphology corrected through targeted fine-tuning

Seed	Token	Gold Label	Baseline Prediction	Targeted Fine-tuned Prediction
41	τῷ	DET Case=Dat Definite=Def Gender=Masc,Neut Number=Sing PronType=Dem	DET Case=Dat Definite=Def Gender=Masc Number=Sing PronType=Dem	DET Case=Dat Definite=Def Gender=Masc,Neut Number=Sing PronType=Dem
41	τοῖσι	DET Case=Dat Definite=Def Gender=Masc,Neut Number=Plur PronType=Dem	DET Case=Dat Definite=Def Gender=Neut Number=Plur PronType=Dem	DET Case=Dat Definite=Def Gender=Masc,Neut Number=Plur PronType=Dem
43	ἄφες	VERB Aspect=Perf Mood=Imp Number=Sing Person=2 Tense=Past VerbForm=Fin Voice=Act	VERB Aspect=Perf Case=Nom Gender=Masc Number=Sing Tense=Past VerbForm=Part Voice=Act	VERB Aspect=Perf Mood=Imp Number=Sing Person=2 Tense=Past VerbForm=Fin Voice=Act

Seed	Token	Gold Label	Baseline Prediction	Targeted Fine-tuned Prediction
43	ἀπόδος	VERB Aspect=Perf Mood=Imp Number=Sing Person=2 Tense=Past VerbForm=Fin Voice=Act	VERB Aspect=Perf Mood=Sub Number=Sing Person=2 Tense=Past VerbForm=Fin Voice=Act	VERB Aspect=Perf Mood=Imp Number=Sing Person=2 Tense=Past VerbForm=Fin Voice=Act
43	τῷ	DET Case=Dat Definite=Def Gender=Masc,Neut Number=Sing PronType=Dem	DET Case=Dat Definite=Def Gender=Masc Number=Sing PronType=Dem	DET Case=Dat Definite=Def Gender=Masc,Neut Number=Sing PronType=Dem

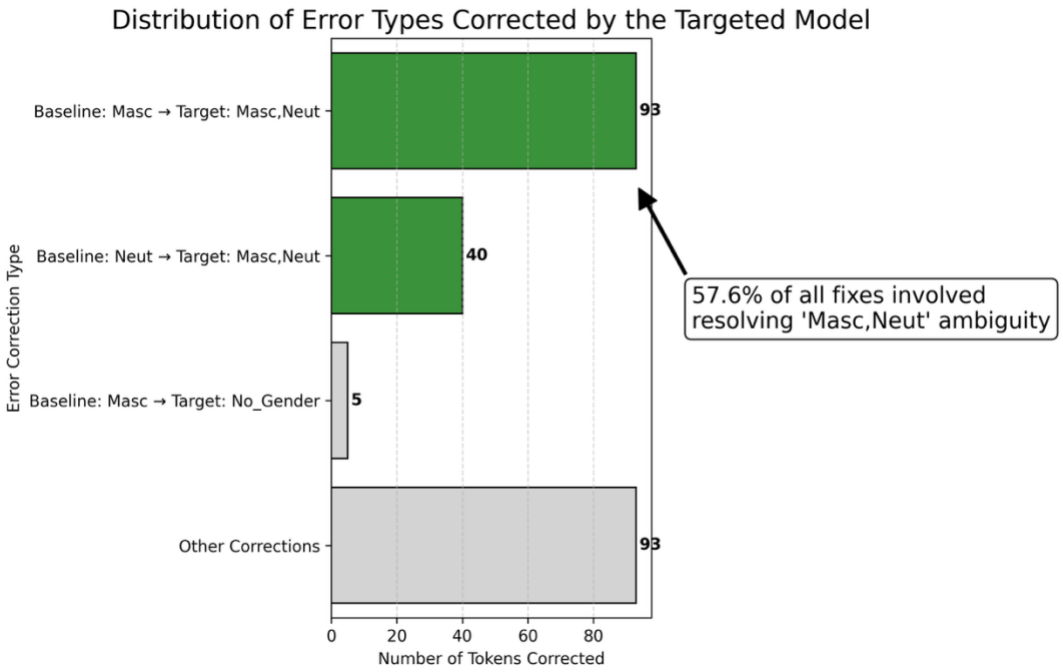

Figure 10 Errors involving Gender=Masc,Neut feature-value pairs composed 57.6% of the resolved errors within the highest performing models across the 5 random seeds.

4.3.4 Statistical Significance Analysis

The Exact Two-sided McNemar's test was applied to compare between the baseline and corresponding targeted fine-tuned model predictions during the evaluation phase for each model. The resulting p-values supported statistical significance for the highest performing models across all 5 random seeds. For seeds 41, 43, 44, and 45, the p-

value was less than 0.001, and seed 42 resulted in a p-value of 0.003. These p-values are less than the standard threshold of $\alpha=0.05$. Therefore, the null hypothesis is rejected indicating that the targeted fine-tuning resulted in statistically significant accuracy gains for the high-risk slice for the highest performing models across all 5 random seeds.

4.3.5 Summary of Support for Hypothesis 2

Hypothesis 2 stated: *Targeted fine-tuning yields significantly higher fine-grained morphological tagging accuracy for the high-risk slice compared to a standard fine-tuned Ancient-Greek-BERT model.* Evidence presented in this section included quantitative improvements, statistical validation, and an error analysis which demonstrated targeted error resolution. Therefore, the results presented in this section provide support for hypothesis 2.

4.4 Sensitivity Analysis

With research question 3 and hypothesis 3 in mind, this section provides results related to the performance of the high-risk slice predictions for various oversampling factors, loss function hyperparameters, and threshold parameters used to define the high-risk slice. The oversampling and loss analysis provides multiple views on the data. It analyzes aggregate performance across all 90 experimental configurations, and it analyzes peak performance for the highest performing models for each of the 5 random seeds. The threshold parameter ablation study utilizes only the highest performing model from the set of 90 experimental configurations.

4.4.1 Oversampling Factor Impact Analysis

Hypothesis 3 proposed that high-risk slice performance is sensitive to the oversampling factor, with intermediate oversampling achieving the highest mean

accuracy improvement among tested factors. Oversampling factors of 2x, 3x, and 4x were utilized during model targeted fine-tuning. As shown in Table 15, when averaged across all 90 model configurations the 3x oversampling factor produced the highest accuracy gains with an average accuracy gain of 2.7 percentage points.

Table 15 Average accuracy changes on the high-risk slice for all 90 model configurations across oversampling factors

Oversampling Factor	Average Accuracy Change (pp)
2x	+0.8
3x	+2.7
4x	+2.3

However, when isolating just the highest performing models for each of the 5 random seeds, the results differ as compared to the average across all 90 models. Table 16 illustrates that a range of 2x, 3x, and 4x oversampling factors yielded optimal results depending on the difficulty of the high-risk slice and other hyperparameter tuning.

Table 16 Oversampling highest performing models accuracy change on the high-risk slice for the models for each random seed across oversampling factors

Seed	Oversampling	Peak Accuracy Gain (pp)
41	3x	+4.5
42	2x	+5.2
43	3x	+6.4
44	4x	+5.7
45	4x	+5.3

4.4.2 Impact of Loss Function Hyperparameters

Whereas oversampling exposed the training process to elements of the high-risk slice at a higher rate than organically found in the training set, the loss functions and associated hyperparameter tuning determine how the model weights are updated in response to the oversampling exposure. The loss functions utilized include class-balanced, class-balanced focal, and cross entropy.

Figure 11 provides an overview of the accuracy change for each model across oversampling factors and loss functions.

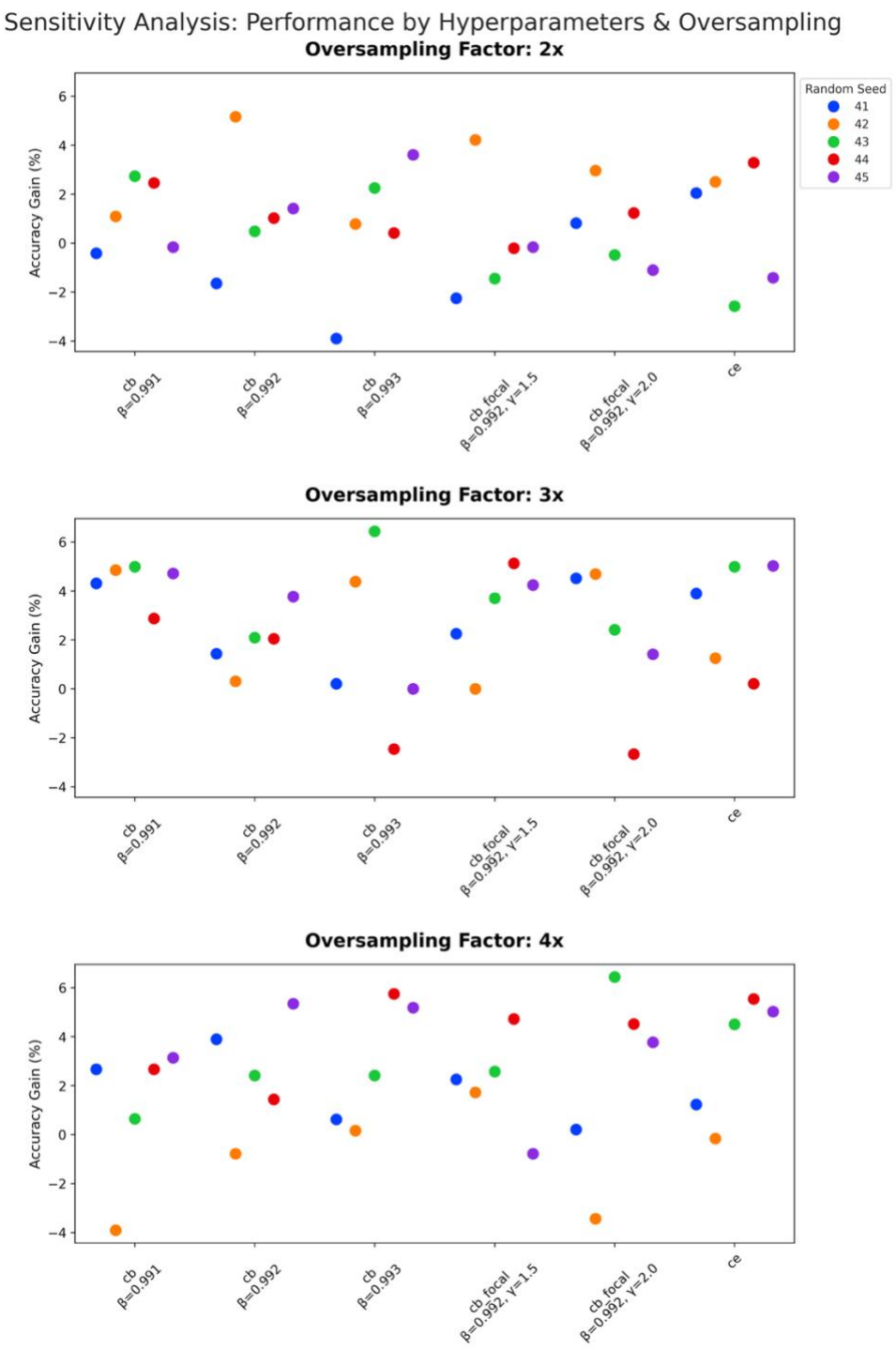

Figure 11 Model performance for various oversampling factors loss functions, and hyperparameters.

As shown in Table 17, in aggregate across all 90 model configurations, Cross Entropy loss exhibited the highest mean accuracy change (+2.4pp). Average accuracy gains were also evident in models trained with class-balanced (+1.9pp) and class-balanced + focal (1.7pp). Cross entropy loss produced the highest mean accuracy gain across the experimental configurations.

Also, due to the various hyperparameters for each model, it should be noted that 15 models were trained with cross-entropy loss, 45 models were trained with class-balanced loss, and 30 models were trained with class-balanced + focal loss. As an experiment design decision, the models trained with class-balanced + focal loss set a constant Beta (β) value of 0.992 to isolate the focusing parameter Gamma γ. Therefore, different denominators were used in calculating the average accuracy change across the loss functions. Additionally, a comparison between models trained with class-balanced loss with a β value of 0.992 exhibited increased accuracy (+1.9pp) as compared to class-balanced + focal loss trained models which exhibited an accuracy of (+1.7pp).

Table 17 Average accuracy changes on the high-risk slice for all 90 model configurations across loss functions

Loss Function	Beta β	Gamma γ	Average Accuracy Change (pp)
Cross Entropy	N/A	N/A	+2.4
Class Balanced	0.991, 0.992, 0.993	N/A	+1.9
Class Balanced (β = 0.992)	0.992	N/A	+1.9
Class Balanced + Focal	0.992	1.5, 2.0	+1.7

Table 18 provides the hyperparameters for loss functions of the highest performing targeted fine-tuned models for each random seed. The results indicate that Class balanced loss (without, and in one case with focal loss) typically produced the

largest increase in accuracy across the high-risk slice for the highest performing models per random seed. β values of 0.992 and 0.993 both exhibited increased accuracies, but a β value of 0.991 did not appear in the highest performing models.

Table 18 Accuracy change on the high-risk slice for the highest performing models for each random seed across loss functions

Seed	Loss Function	Beta (β)	Gamma (γ)	Accuracy Change (pp)
41	Class balanced + Focal	0.992	2.0	+4.5
42	Class balanced	0.992	N/A	+5.2
43	Class balanced	0.993	N/A	+6.4
44	Class balanced	0.993	N/A	+5.7
45	Class balanced	0.992	N/A	+5.3

4.4.3 Threshold Ablation Study

To further evaluate high-risk slice definition thresholds, a systematic ablation study was conducted. The highest performing model configuration (consisting of seed 43, oversampling factor of 3x, and Class-Balanced Loss with a β=0.993) was utilized. The four threshold parameters that were evaluated included the minimum error threshold (τ_e), diversity cap, minimum frequency floor (τ_f), and candidate cap.

The results of the analysis showed that the minimum error threshold (τ_e) and diversity cap had a noticeable impact on the high-risk slice composition. However, the minimum frequency floor (τ_f) and candidate cap did not affect slice composition on the tested ranges (minor performance differences across the runs reflect retraining variances rather than threshold induced slice changes). Lowering the minimum error threshold from 0.05 to 0.03 expanded the high-risk slice size from 621 to 7709 tokens (as shown in Table 19) and lowered the global macro F1

from 0.642 to 0.640 (as shown in Table 20). Raising the minimum error threshold to 0.07 reduced the high-risk slice size from 621 to 334 tokens and reduced global macro F1 to 0.639. These results support a minimum error threshold value of 0.05 as a balanced value among those tested. This value provided a practical compromise which effectively balanced between an overly broad and diluted high-risk slice and a very narrow high-risk slice that reduced evaluation stability.

The initial ablation with the highest performing model suggested that a diversity cap of 0.45 could yield increased performance as compared to a value of 0.35. To obtain a more nuanced understanding of the diversity cap values and resulting performance, an additional multi-seed ablation beyond the single highest performing model was conducted. The five highest performing models were analyzed against diversity cap values of 0.25, 0.35, 0.45, and 0.55. Diversity cap values of 0.25 and 0.35 produced a generally smaller high-risk slice, and 0.45 and 0.55 produced broader high-risk slices with higher mean accuracy.

The multi-seed results (as shown in Table 21) showed that no single diversity-cap value uniformly produced the highest performance across all evaluation configurations. A diversity cap of 0.35 produced the highest mean global accuracy (0.916), while a diversity cap of 0.55 produced the highest mean global Macro F1 score (0.650). Diversity cap values of 0.45 and 0.55 produced the strongest mean high-risk accuracy (0.730 and 0.729), however these gains were also accompanied by much larger average slice sizes compared with the 0.35 configuration. With this result in mind, the 0.35 value was retained for the primary experiment because it was the original pre-ablation design choice used

consistently across all 90 experimental model runs and because the multi-seed analysis showed that it preserved the highest mean global accuracy. Future research utilizing diversity caps of 0.45 or 0.55 may yield additional improvements. Overall, the study indicated that the diversity cap and the minimum error threshold are the most sensitive parameters for slice definition.

Table 19 Threshold Ablation Slice Composition Results
For the highest performing model

Setting	Min Error	Min Frequency	FV Cap	Diversity Cap	Risky FVs	Risky Tokens
Base	0.05	5	2000	0.35	8	621
diversity= 0.25	0.05	5	2000	0.25	8	621
diversity= 0.45	0.05	5	2000	0.45	12	1498
diversity= 0.55	0.05	5	2000	0.55	12	1498
τ_e = 0.03	0.03	5	2000	0.35	21	7709
τ_e = 0.07	0.07	5	2000	0.35	4	334
τ_f = 3	0.05	3	2000	0.35	8	621
τ_f = 10	0.05	10	2000	0.35	8	621
cap = 1000	0.05	5	1000	0.35	8	621
cap = 4000	0.05	5	4000	0.35	8	621

Table 20 Threshold Ablation Performance Results
For the highest performing model

Setting	Global Accuracy	Global Macro F1	High-risk Slice Accuracy
Base	0.916	0.642	0.691
diversity=0.25	0.915	0.645	0.686
diversity=0.45	0.917	0.652	0.768
diversity=0.55	0.918	0.650	0.752
τ_e = 0.03	0.913	0.640	0.879
τ_e = 0.07	0.913	0.639	0.503
τ_f = 3	0.914	0.643	0.676
τ_f = 10	0.912	0.639	0.689
cap = 1000	0.917	0.650	0.712
cap = 4000	0.913	0.652	0.689

Table 21 Threshold Ablation Multi-Seed Diversity Cap Mean Results For the highest performing model per seed

Setting	Risky FVs (mean)	Risky Tokens (mean)	Global Accuracy (mean)	Global Macro F1 (mean)	High-risk Slice Accuracy (mean)
diversity=0.25	7.0	525.6	0.915	0.644	0.613
diversity=0.35	7.4	574.0	0.916	0.638	0.649
diversity=0.45	10.8	1426.6	0.914	0.637	0.730
diversity=0.55	10.8	1426.6	0.915	0.650	0.729

4.4.4 Summary of Sensitivity Analysis and Support for Hypothesis 3

Hypothesis 3 stated: *High-risk slice performance is sensitive to the oversampling factor, with intermediate oversampling achieving the highest mean accuracy improvement among tested factors.* Evidence presented in this section included a quantitative analysis of the impact for various oversampling factors, loss function hyperparameter configurations for both the highest performing models and averaged across all models, and an ablation study of high-risk slice threshold parameters. The average increase in accuracy peaked at a moderate oversampling factor of 3x, with cross-entropy loss achieving the highest mean accuracy gain as compared to class-balanced and class-balanced + focal loss. The highest performing models for each random seed did not exhibit a pattern related to oversampling factors and trended towards largest performance gains with class-balanced loss. Therefore, the results presented in this section provide support for hypothesis 3 in the average case, with the additional insight that maximizing performance requires tuning of oversampling factors and loss function hyperparameters.

4.5 Impact on all Slices

Research question 1 focused on the impact that the targeted fine-tuning process would have across all slices at the global model level. Hypothesis 1 stated: *Targeted fine-tuning does not negatively impact global macro F1 performance compared with a*

standard fine-tuned model. The global F1 of the highest performing models for each random seed was analyzed across the full test set, not just within the high-risk slice. Global macro F1 was selected as the metric to evaluate for significant global regression, as macro F1 is sensitive to unbalanced data with rare classes (Takahashi et al., 2022b). In this case, statistical testing was not applied to evaluate the statistical significance of the global macro F1 change because the purpose is to check for directional regression. Table 22 provides the values between the baseline standard fine-tuned model and the targeted fine-tuned model. Regressions are not observed for any of the models. Rather, macro F1 scores consistently increased. Therefore, hypothesis 1 is supported.

Table 22 Global F1 change for the highest performing models for each seed

Seed	Baseline Model Global Macro F1	Targeted Fine-tuned Model Global Macro F1	Change
41	0.921	0.939	+0.018
42	0.919	0.940	+0.021
43	0.917	0.939	+0.022
44	0.919	0.938	+0.019
45	0.920	0.937	+0.017

4.6 Error Topology and Class Confusion Analysis

To identify the specific morphological features that contributed to accuracy gains or regression, an analysis of error reductions beyond the aggregate provided insight. The analysis identified areas where the targeted fine-tuning had the most significant error reduction impact on specific features. As shown in Figure 12, a delta confusion matrix was developed for the primary morphological features (to include UPOS) within the test dataset. Details of the delta confusion matrix and a deep dive on specific features follows.

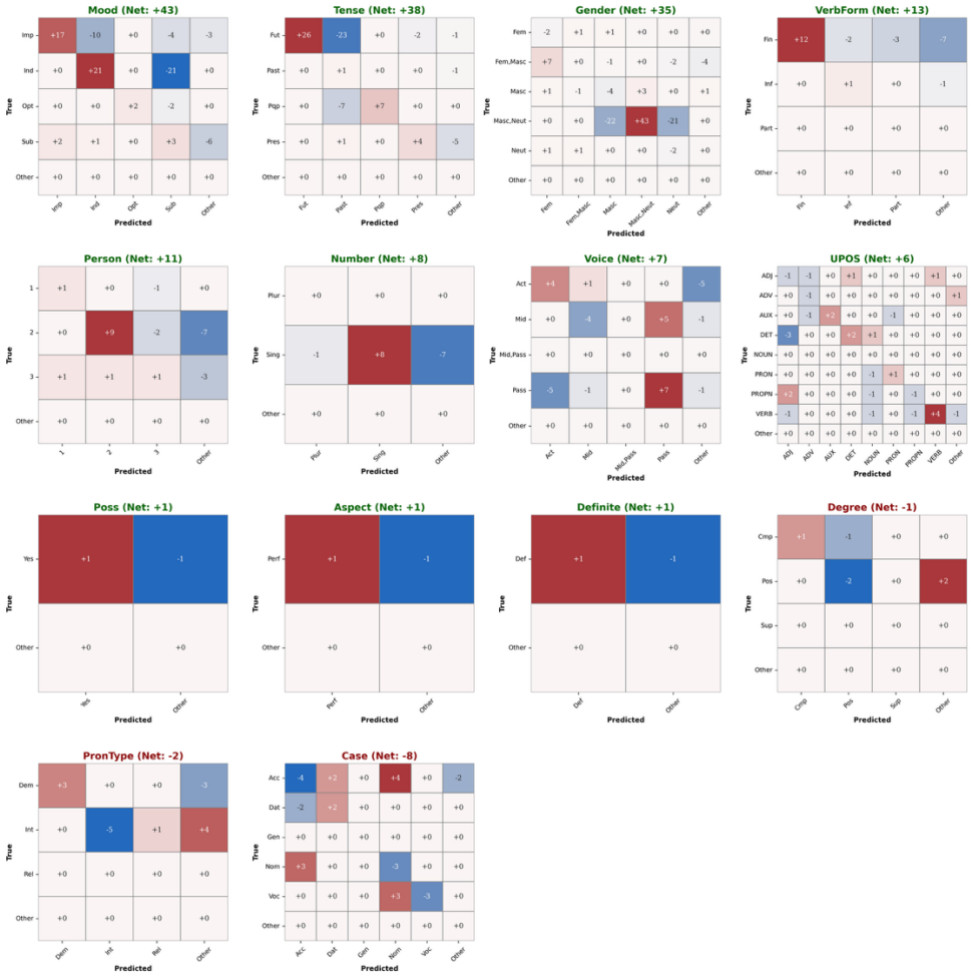

Figure 12 Delta confusion matrices for morphological features. Values show the net change in predictions counts between the baseline and targeted fine-tuned models aggregated across the highest performing model for all five seeds

The delta confusion matrix illustrates where the greatest change occurs within a morphological feature, comparing the confusion matrix produced by the baseline model with the confusion matrix produced by the targeted fine-tuned model (Shi et al., 2025). The interpretation of the delta confusion matrix differs from a standard confusion matrix. The delta confusion matrices use the following convention:

- Red cells (positive values) indicate an increase in prediction count[7]

[7] Note: Shi et al. visualized positive deltas in blue; this research visualizes positive deltas in red for emphasis and readability.

- Blue cells (negative values) indicate a decrease in prediction count
- The counts in each row should have a net sum value of 0
- A positive targeted fine-tuning impact is characterized by red cells along the diagonal (i.e., increased correct predictions) and blue cells not on the diagonal (i.e., decreased errors)

A discussion on all fourteen morphological features would be lengthy. Therefore, a few important select features are presented to illustrate the most significant areas of accuracy change. Both positive and negative illustrative examples are provided for a balanced view on the results.

4.6.1 Systematic Confusion Reduction (Gender)

The Gender feature, as previously discussed in section 4.3.3 Error Correction Analysis, exhibited a high degree of prediction errors. The baseline models repeatedly predicted Gender=Masc or Gender=Neut where the gold labels indicated that the prediction should have been Gender=Masc,Neut. The targeted fine-tuned models corrected this systemic error to a large degree, with an increase of 43 predictions for the feature-value pair Gender=Masc,Neut, a decrease of 21 predictions for the feature-value pair Gender=Masc, and a decrease of 22 predictions for the feature-value pair Gender=Neut. This indicates that the baseline models defaulted to the majority class (i.e., Masc or Neut) for the prediction instead of the minority class (i.e., Masc,Neut). The shift indicates that the targeted fine-tuned models successfully and specifically targeted the area of ambiguity for correction.

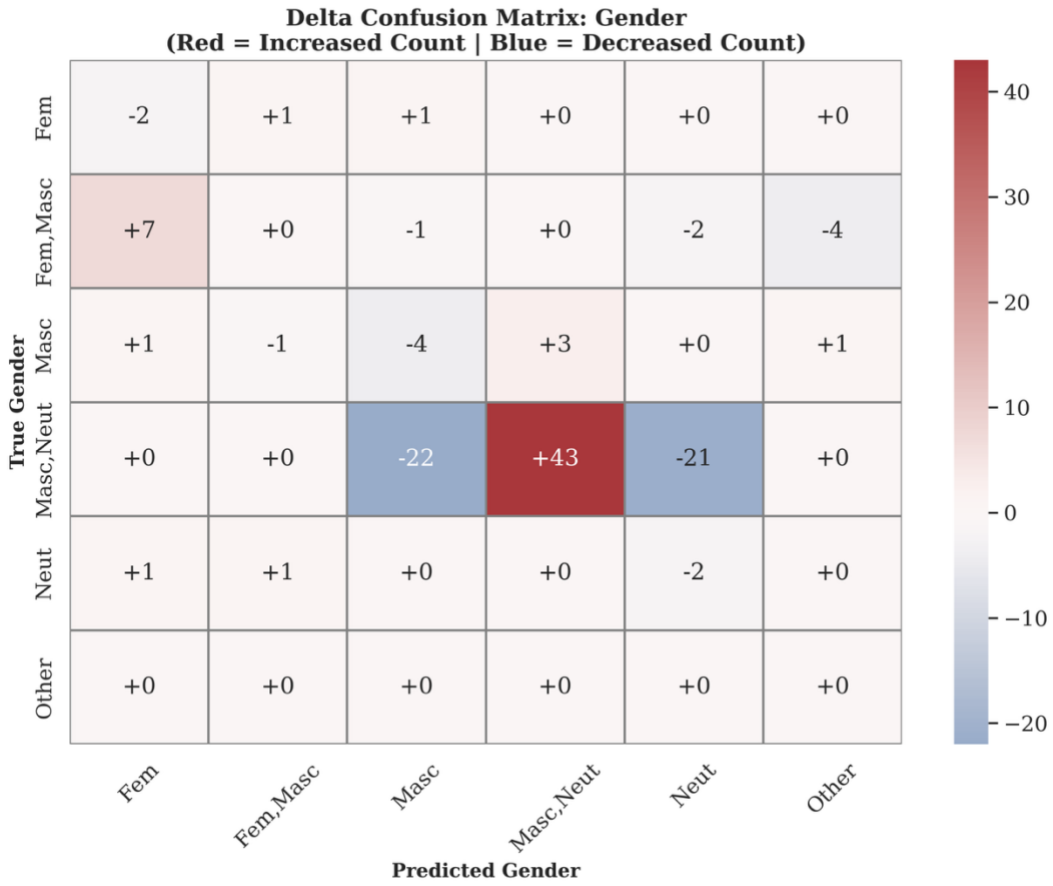

Figure 13 Gender feature delta confusion matrix

4.6.2 Systematic Confusion Reduction (Tense)

Similar to the Gender feature, the Tense feature exhibited prediction errors. As shown in Figure 14, the baseline models repeatedly predicted Tense=Past or Tense=Present where the gold labels indicated that the prediction should have been Tense=Future. The targeted fine-tuned models corrected a total of 26 errors (23 Tense=Past, 2 Tense=Present, 1 Tense=Other). This result indicates that the targeted fine-tuned models successfully corrected morphological errors. Further, the result indicates that the targeted fine-tuning generalizes across features, to include nouns (i.e. Gender) and verbs (i.e. Tense).

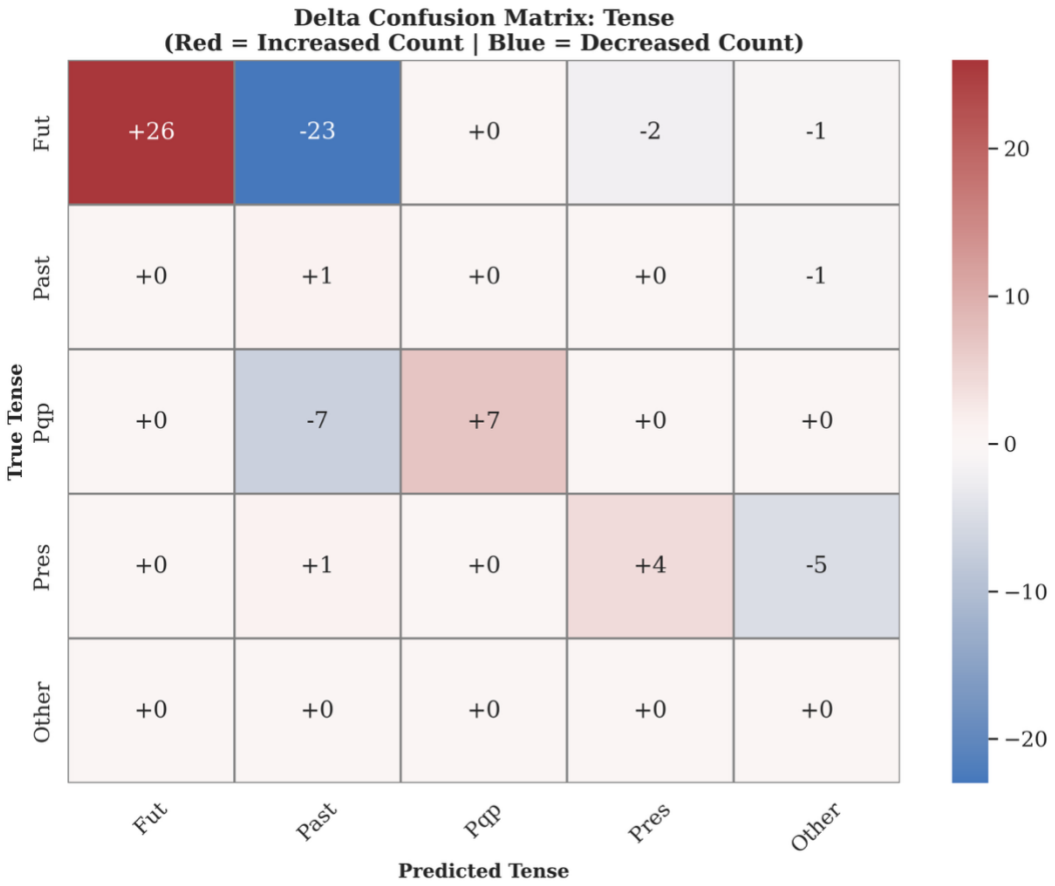

Figure 14 Tense features delta confusion matrix

4.6.3 Confusion Regression (Case)

Regression of certain features also occurred, but on a limited scale. Three of the fourteen features indicated regressions as shown in Figure 12. The Case feature exhibited the largest amount of total regression across the values (illustrated in Figure 15), with a net decrease of 8 correct Case predictions. This indicates that the fine-tuned model did not improve the Case feature accuracy within the slice and slightly worsened it.

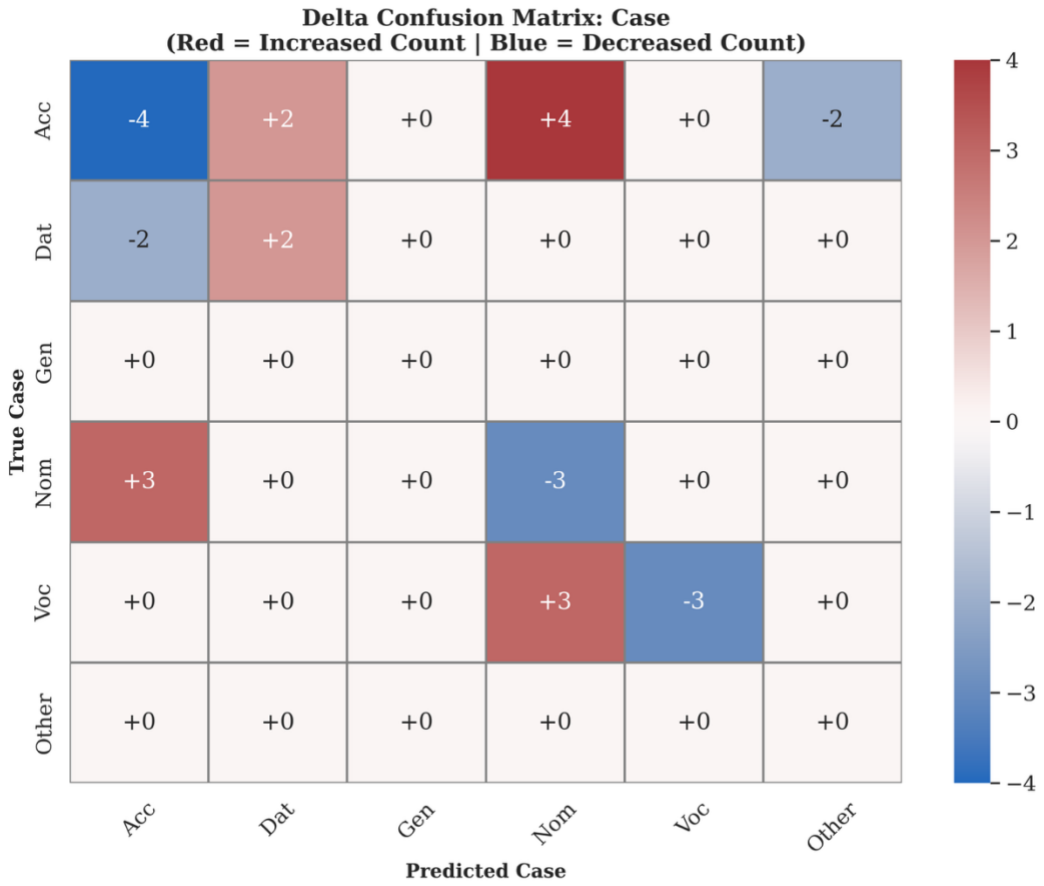

Figure 15 Case feature delta confusion matrix

4.6.4 Feature Stability

The delta confusion matrices also provide evidence for feature stability. As shown in Figure 12, the highest risk features (e.g., Gender, Tense) show accuracy gains through the targeted fine-tuning process with the held-out test data set. Other lower risk features such as UPOS, Person, and Voice exhibit stability following the targeted fine-tuning process. The features that did experience regression were bounded (Case -8 count), indicating that catastrophic forgetting did not occur.

4.7 Training Dynamics

The training loss trajectories for the baseline models compared to the targeted fine-tuned models were also analyzed. As shown in Figure 16, the average of the 90

targeted fine-tuned models reached a training loss near 0.2 after 1 epoch, whereas the baseline models' training loss is approximately 1.5 after 1 epoch. After 2 epochs, the targeted fine-tuned models stabilize with a training loss of approximately 0.1. Baseline models show decreased training loss even after 5 epochs.

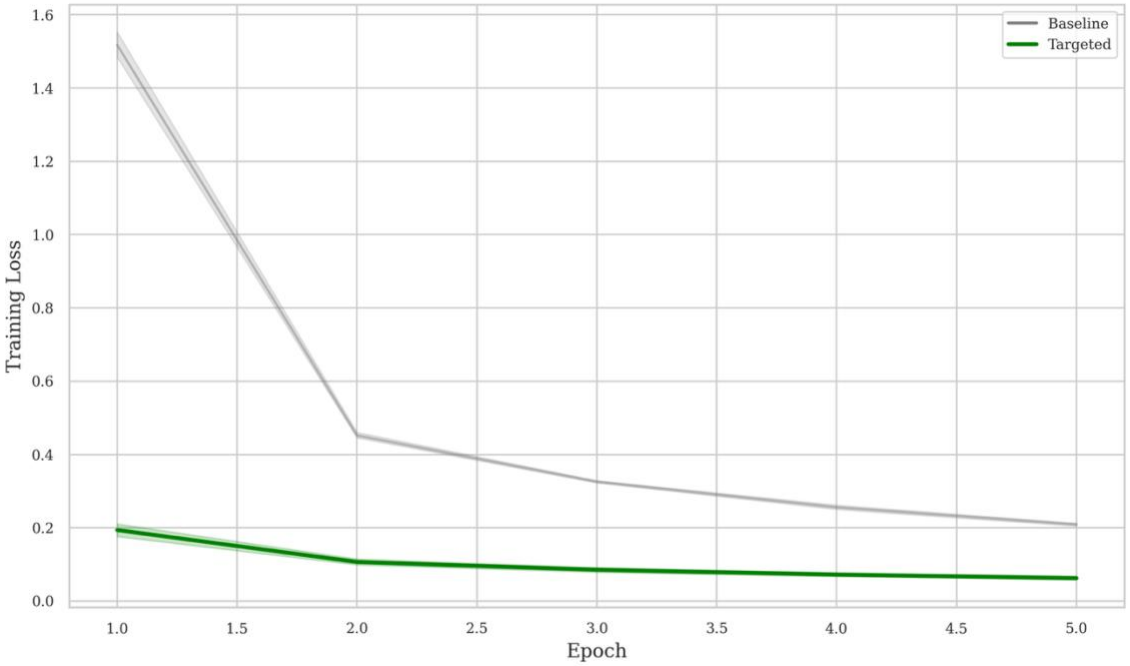

Figure 16 Comparison of training loss across epochs for baseline and the targeted fine-tuned models

However, it must be noted that the baseline and targeted models are initialized from different states. The baseline models are initialized from the Ancient-Greek-BERT model. The targeted models are initialized from the baseline standard fine-tuned model for their respective seed, as depicted in Figure 17. For example, if the baseline model was initialized with seed 42, then the targeted model would initialize with the baseline model that was also initialized with seed 42.

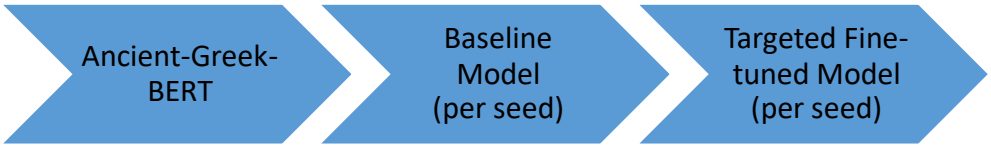

Figure 17 Model Initialization Flow.
The baseline model is initialized with the Ancient-Greek-BERT model and a selected random seed. The targeted fine-tuned model is initialized with the baseline model of the same corresponding random seed as selected for the targeted fine-tuned model.

Given that baseline and targeted models are not initializing from the same weighting state, the training loss curves for speed of optimization are not directly comparable. However, Figure 16 shows that targeted fine-tuning reaches a low level of training loss quickly and stabilizes after two epochs. Further analysis of a high performing targeted fine-tuning run (seed=43, oversampling=3x, class-balanced loss, β=0.993) shows that overall validation set accuracy increased as shown in Table 23. These results indicate that the accuracy gains are most pronounced in the first two epochs, with further epochs showing diminished increases in accuracy gains. This behavior suggests that targeted fine-tuning can be achieved with a modest number of training epochs (i.e., 2 or 3) compared with a typical 5 epochs of training. Therefore, a potential exists for an approximate 40-60% reduction of training compute costs per model.

Table 23 Validation Accuracy Change per Epoch for the highest performing model

Epoch	Accuracy
1	90.73%
2	91.55%
3	92.05%
4	92.10%
5	92.11%

4.8 Downstream Translation Results

A small-scale pilot was conducted to evaluate whether the morphological tag accuracy gain from DART-FT yielded a statistically significant improvement in translation quality compared with the baseline model tags. On the primary metric of chrF++, DART-FT improved by +0.24 (p=0.0408). Translation Edit Rate (TER) also improved as shown in a decrease from 59.42 to 59.27 (-0.15). Further, in the subset where the baseline and DART-FT tag sequences differed (N=232), DART-FT achieved a +0.37 chrF++ gain and a -0.23 reduction in TER.

The overall direction of translation quality in the limited scope pilot was positive, but some limitations should be noted. The decoder altered the Latin it generated as output in 83 of the 232 targeted instances (35.8%). This suggests that the translator relied heavily on its pre-training and disregarded some of the newly introduced morphological information. Overall, the pilot should be approached as limited evidence and not as a full-scale detailed translation study.

4.9 Summary

Empirical evidence was provided in this chapter based on the experiments conducted utilizing the methodology stated in Chapter 3. Hypotheses 1 and 2 were supported and demonstrated that *a targeted fine-tuned Ancient-Greek-BERT model*

reduces mislabeling in fine-grained Ancient Greek morphological tagging. Hypothesis 3 was also supported regarding that specific hyperparameter tuning is necessary to achieve peak performance. Overall, the targeted fine-tuning method is effective.

Chapter 5—Discussion and Conclusions

5.1 Discussion

DART-FT addressed a low resource language natural language processing challenge. Fine-grained morphological tagging can be especially error prone for tokens associated with ambiguous or sparse feature patterns. Transformer models (i.e., Ancient-Greek-BERT) can achieve a high degree of aggregate accuracy but fail to model correct morphology for rare or morphologically complex words typically due to data sparsity.

This study introduced a methodology designed to identify and repair the specific error-prone areas of weakness within the model. Previous research primarily focused on global metrics (i.e., global accuracy/Macro F1) and lacked an exploration of a targeted and precise method to increase slice level metrics for error prone regions. The DART-FT pipeline addressed the gap by introducing a targeted fine-tuning method. The method utilized a multi-pass pipeline to identify and isolate a high-risk slice which represented a subset of data where the model exhibited weakness. Sections 5.1.1 through 5.1.6 interpret these findings in relation to RQ1, RQ2, and RQ3.

Targeted fine-tuning produced statistically significant accuracy gains on the high-risk slice. The accuracy gains did not come at the expense of global accuracy. This indicates that the method was effective while not inducing catastrophic forgetting.

5.1.1 Error Reduction in Gender Feature Syncretism

In section 4.3.3 Error Correction Analysis showed 57.6% of corrected predictions in the highest performing targeted fine-tuned models involved the Gender=Masc,Neut

feature-value pair. This relatively high concentration indicates that the baseline models exhibit a concentrated area of failure which the targeted fine-tuning method resolved.

The observed concentration of failures appears to be primarily driven by the structural class imbalance within the PROIEL treebank combined with syncretism (tokens that appear the same across multiple morphological features). Ancient Greek exhibits syncretism as shown in Table 1, where tokens can appear identical across different features such as Gender, Case, and Number without the associated lexical context. Baseline models rely on the surrounding contextual cues to support Gender determination.

Additionally, an analysis of the PROIEL treebank reveals structural imbalance of the Gender feature which causes issues with contextual learning. For example, there are 8 part-of-speech tags that utilize the Gender feature. An aggregation of the data reveals that 54.4% of the tokens with a Gender feature are tagged as Gender=Masc, 17.7% as Gender=Neut, and only 2.3% are tagged with the combined Gender=Masc,Neut label. (*Statistics of Gender in UD_Ancient_Greek-PROIEL*, 2024)

Cross-entropy (CE) training handles this context-dependent ambiguity by selecting the most common label. As shown in Figure 13, the baseline model defaults to the majority classes (Masc or Neut) rather than correctly selecting Masc,Neut. Targeted fine-tuning trains the model to increase learning on the underrepresented tokens, which increases the weight of the gradient contributions for Gender=Masc,Neut through penalizing the model for majority class guesses. Ultimately, this suggests that the targeted fine-tuning methodology enables post-training intervention for systematic error patterns

on imbalanced data and is likely to be substantially less computationally expensive than approaches that require additional pre-training.

5.1.2 CE and Class Balanced Loss Robustness vs. Peak Performance

As shown in Table 17, CE loss produced the highest mean high-risk slice accuracy gain across all 90 model configurations (+2.4 percentage points), compared with Class Balanced (CB) loss and Class Balanced Focal (CB+F) loss which exhibited lower accuracy changes. Similarly, Table 15 shows that 3x oversampling achieved the highest average accuracy gain across the tested oversampling factors.

However, the top performing models for each seed were trained with CB or CB+F loss and not CE loss, as shown in Table 18. Further, Table 16 shows the oversampling factor associated with peak performance varied by seed (i.e., 2x-4x). This pattern suggests a tension between robustness and peak performance. CE loss is comparatively robust because it applies a per-example weighting objective, does not reweight classes by frequency, and requires minimal hyperparameter tuning to be effective at increasing accuracy. It acts as a stable but imprecise tool which is robust across multiple different configurations.

In contrast, CB and CB+F loss introduce hyperparameters that determine the level of emphasis of the rare classes. When these hyperparameters are aligned with the specific imbalance structure of the high-risk slice composition, the accuracy gains can exceed the accuracy gains from CE loss. However, when the CB and CB+F loss hyperparameter alignment is suboptimal, CE loss can produce superior accuracy gains. Similarly, the oversampling factor of 3x performs well on average, but the oversampling factor that produces peak accuracy gains varies across random seeds.

Practically, these results suggest that when training budgets are limited (in terms of time and/or money), CE loss is appropriate as a default loss function with an oversampling factor of 3x. When tuning budget is available to maximize slice-level performance, CB loss or CB+F loss are appropriate with a bounded search over hyperparameters and oversampling factors.

5.1.3 Multi-label Error Regression in Case and PronType Features

In aggregate the features within the highest performing targeted fine-tuned models for each seed showed accuracy increases. However, as shown in Figure 15, the Case (-8), PronType (-2), and Degree (-1) features exhibited a net decrease of correct predictions. The regression of the Degree feature represents a small regression of a single correct prediction, the Case and PronType regressions are more significant.

It is noteworthy that the feature-value pairs Case=Voc and PronType=Int were included in the high-risk slice across all 5 random seeds (Table 24) and still exhibited regression of correct predictions. Conversely, other feature-value pairs such as Gender=Masc,Neut (+35) and Poss=Yes (+1) were also targeted across all 5 seeds within the high-risk slice and experienced gains in correct predictions. This suggests cross-feature interaction (negative transfer) occurred.

For the Case feature, the net loss of correct predictions was concentrated among the Case=Nom (-3), Case=Acc (-4), and Case=Voc (-3) values, and for the PronType feature, the new errors in the PronType=Int value (-5) slightly exceeded the gains in the PronType=Dem (+4). As shown in Table 1, interrogative pronouns (e.g., τίνα) can exhibit multiple ambiguities spread across multiple feature-value pairs. The model must rely on context to assign the correct tags, which may couple PronType and Case feature errors

within the larger morphological bundle. These regressions suggest that in cases of highly complex ambiguous fine-grained morphology, the targeted fine-tuning intervention may introduce localized tradeoffs across tightly coupled features rather than uniformly improving all features within the bundle. To mitigate the impacts of negative transfer, future work should consider methods including alternative architectural and training approaches. Potential approaches could include implementation and evaluation of feature-specific loss weighting (Kendall et al., 2018) or other task specific optimization strategies to reduce the risk of optimization conflicts. Additionally, evaluating whether intermediate-task transfer learning (Pruksachatkun et al., 2020) could be adapted to mitigate regressions, could yield a beneficial approach.

5.1.4 Global Stability and Catastrophic Forgetting Avoidance

The potential for catastrophic forgetting exists when oversampling introduces multiple copies of the same data into the training pipeline. However, the aggregate mean improvement of +0.0194 in global macro F1 score (Table 22), indicates no evidence of global regression which would be consistent with catastrophic forgetting. There are two factors that likely contributed to the global stability.

First, the oversampling process maintained the sentence level structure, even though the high-risk slice was targeting specific tokens. The retained context acted to anchor the model. Second, a relatively small number of training epochs were needed to stabilize the model as noted in Section 4.7 Training Dynamics. The small number of epochs limited the potential for overfitting on specific high-risk tokens. Overall, this enabled the model to repair the high-risk slice without damaging the overall performance.

5.1.5 Applicability and Generalizability Boundaries

As noted in Section 1.8 Research Limitations, the DART-FT methodology has several prerequisites to achieve successful outcomes. If this method is applied to other low-resource languages or to other domains, the following must be available:

- A gold labeled dataset (e.g., treebank) with token-level annotations: This is necessary for model evaluation and high-risk slice identification.
- A pre-trained model: This is necessary as a starting point for the methodology to conduct post-training.
- A defined and finite decomposable feature set: The set of features must be finite and decomposable to define the high-risk slice.

Therefore, while only validated on Ancient Greek, the prerequisites of the DART-FT method suggest it may be suitable for low (but not zero) resource languages or domains.

Beyond these prerequisites, the improvements in tagging accuracy which DART-FT exhibited were dependent on properties of Ancient Greek and the UD PROIEL dataset. As previously mentioned, Ancient Greek exhibits rich morphology and systematic ambiguity which concentrates error patterns into identifiable slices. DART-FT leverages this pattern to identify and correct errors. Other domains with randomly distributed errors may see limits to the benefits which DART-FT provides. Also, PROIEL contains high-quality human annotated labels. Other domains that lack high-quality labels may require additional approaches to ensure label data quality. Therefore, application of DART-FT into domains such as cybersecurity log analysis (as mentioned in the future work section) must take these requirements into account.

Overall, DART-FT provides a targeted and resource efficient potential alternative to pre-training and standard CE-loss fine-tuning by enabling targeted repairs to error-prone areas of models that have limited labeled corpora.

5.1.6 Translation Pilot Implications

The primary contribution was improvement of tagging accuracy on the high-risk slice. Ideally, improved tagging accuracy should have an associated positive downstream impact. The results of the translation pilot provided evidence for marginally, but statistically significant improved translation quality (chrF++) and translation edit rate (TER) compared to the baseline. Although the pilot was not intended to yield a production quality translation pipeline, the results suggest that human translation editors could experience a potential reduction in labor in post-translation which supports the initial motivation.

5.2 Conclusions

The following conclusions are drawn from the research questions and supporting evidence provided in Chapter 4—Results.

5.2.1 Research Question 1 Evaluation and Conclusion

Research Question 1: *Does targeted fine-tuning negatively impact global macro F1 performance compared with a standard fine-tuned model?*

Hypothesis 1: *Targeted fine-tuning does not negatively impact global macro F1 performance compared with a standard fine-tuned model.*

Evaluation and Conclusion: The results support the hypothesis. As shown in Table 22, global F1 scores across the 5 random seeds exhibited no observed decrease and instead showed increases (mean improvement of +0.0194), following targeted fine-

tuning. This confirms that when taking a global view across the entire dataset, targeted fine-tuning did not negatively impact fine-grained tagging F1 score compared with a standard fine-tuned model.

5.2.2 Research Question 2 Evaluation and Conclusion

Research Question 2: *Does targeted fine-tuning increase fine-grained morphological tagging accuracy for the high-risk slice compared to a standard fine-tuned Ancient-Greek-BERT model?*

Hypothesis 2: *Targeted fine-tuning yields significantly higher fine-grained morphological tagging accuracy for the high-risk slice compared to a standard fine-tuned Ancient-Greek-BERT model.*

Evaluation and Conclusion: The hypothesis is strongly supported. Targeted fine-tuning yields significantly higher fine-grained morphological tagging accuracy across the 5 random seeds, with a peak accuracy increase of 6.4 percentage points in the best performing model (Table 13 and Figure 9). The delta confusion matrices (Figure 12) demonstrate the targeted fine-tuning corrected prediction errors in 11 of 14 morphological features such as Gender (Figure 13) and Tense (Figure 14). Statistical testing, utilizing the McNemar test rejected the null hypothesis, with p-values < 0.001 (Table 13). This result indicates that accuracy increase is not a result of random variance.

5.2.3 Research Question 3 Evaluation and Conclusion

Research Question 3: *How sensitive is targeted fine-tuning performance to the degree of oversampling of high-risk tokens?*

Hypothesis 3: *High-risk slice performance is sensitive to the oversampling factor, with intermediate oversampling achieving the highest mean accuracy improvement among tested factors.*

Evaluation and Conclusion: The hypothesis is supported in the average case. On average, a moderate oversampling factor of 3x yields the greatest improvement in high-risk slice accuracy with an average accuracy increase of 2.7 percentage points (Table 15) across the 5 random seeds. However, the highest performing models for each random seed did not exhibit a pattern supporting a moderate oversampling factor of 3x as ideal (Table 16). Catastrophic forgetting did not occur for the peak performing models for each random seed, as demonstrated by the consistent accuracy gains in the morphological analysis (Figure 12).

5.3 Contributions to the Body of Knowledge

This research provides several contributions to the body of knowledge:

1. A high-risk slice identification method. A novel data-driven pipeline for identifying an empirically defined high-risk slice associated with frequent and systematic fine-grained morphological tagging errors was developed. Unlike a static list, the method provides a dynamic high-risk slice which provides the specific error patterns that the targeted fine-tuning method utilizes for error reduction.
2. A statistically validated targeted fine-tuning method. A reproducible method to improve accuracy on the high-risk slice was demonstrated. Statistical analysis validated the mechanism for improving accuracy of the high-risk slice.

5.4 Recommendations for Future Research

Several areas are ripe for potential future research:

1. Iterative applications of DART-FT: The approach was to define the high-risk slice and then perform targeted fine-tuning. As demonstrated, this approach increased accuracy on the high-risk slice. However, what would be the impact if the model output of the targeted fine-tuning process becomes the input for another iteration of high-risk slice identification and targeted fine-tuning? Research into the effects of an iterative approach of DART-FT would potentially provide valuable insights in iterative error reduction.
2. High-risk slice threshold values ablation: The threshold values which support identification of the high-risk slice were selected as design choices. However, future work conducting a full empirical optimization study into the threshold values could yield further insights. Specifically, recommendations to practitioners regarding guidelines for threshold values depending on the use-case and data characteristics could be helpful.
3. Mitigation of negative transfer in tightly coupled features: As discussed in Section 5.1.3 Multi-label Error Regression in Case and PronType Features, future research should explore mitigation strategies to reduce morphological feature-value pair accuracy regressions. Potential areas for exploration could include feature specific loss weighting (Kendall et al., 2018) and intermediate task transfer learning (Pruksachatkun et al., 2020).

These approaches may reduce the risk of cross-feature regressions within tightly coupled morphological features.

4. Alternative dataset/treebank: Experiments with different baseline datasets may yield beneficial results. Datasets such as the Ancient Greek Dependency Treebank (AGDT) (Celano et al., 2021) could be considered as it is similar to the PROIEL Ancient Greek treebank. Specifically, evaluating the change in accuracy and F1 score with a different dataset would be informative.
5. Alternative domains: An exploration of applications outside of the Ancient Greek NLP domain could yield impact. For example, exploring whether a transformer model could be trained with the targeted fine-tuning method to predict rare cybersecurity events within log files, could provide a high impact area of practical application. Alternative definitions of the high-risk slice, such as "Slice Finder" which utilizes an unlabeled data set (Chung et al., 2019), could be explored to identify the focus area for targeted fine-tuning. Further research in this area could enable more general application of the targeted fine-tuning method without the current limitations as described in Section 5.1.5.

References

2024 Global Scripture Access. (2025, August 10). *Wycliffe Global Alliance*. https://www.wycliffe.net/resources/statistics/

Abudouwaili, G., Abiderexiti, K., Shen, Y., & Wumaier, A. (2022). Research on the Uyghur morphological segmentation model with an attention mechanism. *Connection Science*, *34*(1), 2577–2596. https://doi.org/10.1080/09540091.2022.2134843

Adhikari, B., Peltomäki, J., Germi, S. B., Rahtu, E., & Huttunen, H. (2021). Effect of Label Noise on Robustness of Deep Neural Network Object Detectors. In I. Habli, M. Sujan, S. Gerasimou, E. Schoitsch, & F. Bitsch (Eds.), *Computer Safety, Reliability, and Security. SAFECOMP 2021 Workshops* (Vol. 12853, pp. 239–250). Springer International Publishing. https://doi.org/10.1007/978-3-030-83906-2_19

Agresti, A. (2019). An Introduction to Categorical Data Analysis. In *An Introduction to Categorical Data Analysis* (3rd ed., pp. 227–250). John Wiley & Sons, Inc.

Agresti, A., & Min, Y. (2005). Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine*, *24*(5), 729–740. https://doi.org/10.1002/sim.1781

Aitken, J. K., & Dhont, M. (2023). The Septuagint within the History of Greek: An Introduction. *Journal for the Study of Judaism*, *54*(432–449). https://doi.org/10.1163/15700631-bja10073

Allan, R. J. (2014). Changing the Topic: Topic Position in Ancient Greek Word Order. *Mnemosyne*, *67*(2), 181–213.

ALP 2025. (2025, August 10). https://www.ancientnlp.com/alp2025/

Bamman, D., & Crane, G. (2011). The Ancient Greek and Latin Dependency Treebanks. In C. Sporleder, A. van den Bosch, & K. Zervanou (Eds.), *Language Technology for Cultural Heritage* (pp. 79–98). Springer. https://doi.org/10.1007/978-3-642-20227-8_5

Bamman, D., Mambrini, F., & Crane, G. (2009). *An Ownership Model of Annotation: The Ancient Greek Dependency Treebank*. ITA. https://publicatt.unicatt.it/handle/10807/132855

Belinkov, Y., & Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72. https://doi.org/10.1162/tacl_a_00254

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Bentein, K. (2016). *Verbal Periphrasis in Ancient Greek: Have- and Be- Constructions*. Oxford University Press.

Celano, G. G. A., Crane, G., & Almas, B. (2021, January 21). *The Ancient Greek and Latin Dependency Treebank by PerseusDL*. https://perseusdl.github.io/treebank_data/

Celis, E., Keswani, V., Straszak, D., Deshpande, A., Kathuria, T., & Vishnoi, N. (2018). Fair and Diverse DPP-Based Data Summarization. *Proceedings of the 35th International Conference on Machine Learning*, 716–725. https://proceedings.mlr.press/v80/celis18a.html

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chiruzzo, L., Ritter, A., & Wang, L. (Eds.). (2025). *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.0

Codex Sinaiticus—See The Manuscript | John 1:1-3, Folio 247, British Library. (2025, August 9). Codex Sinaiticus Project. https://www.codexsinaiticus.org/en/manuscript.aspx?__VIEWSTATEGENERATOR=01FB804F&book=36&lid=en&side=r&zoomSlider=0#36-1-3-9

CoNLL-U Format. (2025, August 29). https://universaldependencies.org/format.html

Connor, R., Dearle, A., Claydon, B., & Vadicamo, L. (2024). Correlations of Cross-Entropy Loss in Machine Learning. *Entropy*, *26*(6), 491. https://doi.org/10.3390/e26060491

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-Balanced Loss Based on Effective Number of Samples. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9260–9269. https://doi.org/10.1109/CVPR.2019.00949

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Devlin, J. T., Jamison, H. L., Matthews, P. M., & Gonnerman, L. M. (2004). Morphology and the internal structure of words. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(41), 14984–14988. https://doi.org/10.1073/pnas.0403766101

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1923. (1119864). https://doi.org/10.1162/089976698300017197

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O. E., & Jøhndal, M. (2018). The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, *52*(1), 29–65. https://doi.org/10.1007/s10579-017-9388-5

Fagerland, M. W., Lydersen, S., & Laake, P. (2013). The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, *13*(1), 1–8. (89699368). https://doi.org/10.1186/1471-2288-13-91

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.),

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6894–6910). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.552

Gorman, V. B. (2020). Dependency Treebanks of Ancient Greek Prose. *Journal of Open Humanities Data*, *6*(1). https://doi.org/10.5334/johd.13

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Heilmann, J. (2020). Reading Early New Testament Manuscripts: Scriptio continua, "Reading Aids", and Other Characteristic Features. In A. Krauß, J. Leipziger, & F. Schücking-Jungblut (Eds.), *Material Aspects of Reading in Ancient and Medieval Cultures: Materiality, Presence and Performance* (pp. 177–196). De Gruyter. https://www.degruyterbrill.com/document/doi/10.1515/9783110639247-011/html

Helsinki-NLP/opus-mt-tc-bible-big-mul-mul · Hugging Face. (2024, October 9). https://huggingface.co/Helsinki-NLP/opus-mt-tc-bible-big-mul-mul

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18*, 3207–3214.

Herbert Weir Smyth, A Greek Grammar for Colleges, chapter 334. (1920). https://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0007%3Asmythp%3D334

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735. (9710215021). https://doi.org/10.1162/neco.1997.9.8.1735

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 328–339). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021, October 6). *LoRA: Low-Rank Adaptation of Large Language Models*. International Conference on Learning Representations. https://openreview.net/forum?id=nZeVKeeFYf9

Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging* (arXiv:1508.01991). arXiv. https://doi.org/10.48550/arXiv.1508.01991

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 245–251. https://doi.org/10.1109/ACII.2013.47

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Keersmaekers, A., & Mercelis, W. (2024). Adapting transformer models to morphological tagging of two highly inflectional languages: A case study on Ancient Greek and Latin. In J. Pavlopoulos, T. Sommerschield, Y. Assael, S. Gordin, K. Cho, M. Passarotti, R. Sprugnoli, Y. Liu, B. Li, & A. Anderson (Eds.), *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)* (pp. 165–176). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.ml4al-1.17

Kendall, A., Gal, Y., & Cipolla, R. (2018). *Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics*. Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.17863/CAM.25486

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. In D. Lin & D. Wu (Eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 388–395). Association for Computational Linguistics. https://aclanthology.org/W04-3250/

Kondratyuk, D., & Straka, M. (2019). 75 Languages, 1 Model: Parsing Universal Dependencies Universally. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2779–2795). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1279

Kostkan, J., Kardos, M., Mortensen, J. P. B., & Nielbo, K. L. (2023). OdyCy – A general-purpose NLP pipeline for Ancient Greek. In S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.), *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 128–134). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.latechclfl-1.14

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language*, *6*(3), 225–242. https://doi.org/10.1016/0885-2308(92)90019-Z

Liddell, F. D. K. (1983). Simplified Exact Analysis of Case-Referent Studies: Matched Pairs; Dichotomous Exposure. *Journal of Epidemiology and Community Health (1979-)*, *37*(1), 82–84.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007. https://doi.org/10.1109/ICCV.2017.324

Liu, Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., & Finn, C. (2021). Just Train Twice: Improving Group Robustness without Training Group Information. *Proceedings of the 38th International Conference on Machine Learning*, 6781–6792. https://proceedings.mlr.press/v139/liu21f.html

Liu, J., Duan, X., Zhang, R., Sun, Y., Guan, L., & Lin, B. (2021). Relation classification via BERT with piecewise convolution and focal loss. *PLOS ONE*, *16*(9), e0257092. https://doi.org/10.1371/journal.pone.0257092

Liu, Z., & Prud'hommeaux, E. (2022). Data-driven Model Generalizability in Crosslinguistic Low-resource Morphological Segmentation. *Transactions of the Association for Computational Linguistics*, *10*, 393–413. https://doi.org/10.1162/tacl_a_00467

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

McNemar, Q. (1947). Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/https://doi.org/10.1007/BF02295996

Merialdo, B. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics*, *20*(2), 155–171.

MorphGNT. (2025, August 14). http://morphgnt.org/

Moumoulidou, Z., McGregor, A., & Meliou, A. (2021). Diverse Data Selection under Fairness Constraints [Application/pdf]. *LIPIcs, Volume 186, ICDT 2021*, *186*, 13:1-13:25. https://doi.org/10.4230/LIPICS.ICDT.2021.13

Mueller, T., Schmid, H., & Schütze, H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 322–332). Association for Computational Linguistics. https://aclanthology.org/D13-1032/

Nigatu, H. H., Tonja, A. L., Rosman, B., Solorio, T., & Choudhury, M. (2024). The Zeno's Paradox of `Low-Resource' Languages. In Y. Al-Onaizan, M. Bansal, &

Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17753–17774). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.983

Pakray, P., Gelbukh, A., & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, *31*(2), 183–197. https://doi.org/10.1017/nlp.2024.33

Pava, J. N., Meinhardt, C., Badi Uz Zaman, H., Friedman, T., Truong, S. T., Zhang, D., Cryst, E., Marivate, V., & Koyejo, S. (2025, April 22). *Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts | Stanford HAI*. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/policy/mind-the-language-gap-mapping-the-challenges-of-llm-development-in-low-resource-language-contexts

Plank, B., Søgaard, A., & Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 412–418). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2067

Popović, M. (2017). chrF++: Words helping character n-grams. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, & J. Kreutzer (Eds.), *Proceedings of the Second Conference on Machine Translation* (pp. 612–618). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4770

Prior, A., Wintner, S., Macwhinney, B., & Lavie, A. (2010). Translation ambiguity in and out of context. *Applied Psycholinguistics*, *32*(1), 93–111. https://doi.org/10.1017/s0142716410000305

Pruksachatkun, Y., Phang, J., Liu, H., Htut, P. M., Zhang, X., Pang, R. Y., Vania, C., Kann, K., & Bowman, S. R. (2020). Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5231–5247). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.467

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. https://doi.org/10.1109/5.18626

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Rapacz, M., & Smywiński-Pohl, A. (2025). Low-Resource Interlinear Translation: Morphology-Enhanced Neural Models for Ancient Greek. In H. Hettiarachchi, T. Ranasinghe, P. Rayson, R. Mitkov, M. Gaber, D. Premasiri, F. A. Tan, & L. Uyangodage (Eds.), *Proceedings of the First Workshop on Language Models for Low-Resource Languages* (pp. 145–165). Association for Computational Linguistics. https://aclanthology.org/2025.loreslm-1.11/

Reimers, N., & Gurevych, I. (2017). Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In M. Palmer, R.

Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 338–348). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1035

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4902–4912). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.442

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Seeker, W., & Kuhn, J. (2013). Morphological and syntactic case in statistical dependency parsing. *Comput. Linguist.*, *39*(1), 23–55. https://doi.org/10.1162/COLI_a_00134

Shi, H., Norton, L., Ridding, L., Rolph, S., August, T., Wood, C. M., Qie, L., Bosilj, P., & Brown, J. M. (2025). *Habitat Classification from Ground-Level Imagery Using Deep Neural Networks* (arXiv:2507.04017). arXiv. https://doi.org/10.48550/arXiv.2507.04017

Singh, P., Rutten, G., & Lefever, E. (2021). A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek. In S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.), *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage,*

Social Sciences, Humanities and Literature (pp. 128–137). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.latechclfl-1.15

Smyth, H. W. (1920). *A Greek Grammar for Colleges* (1st ed.). American Book Company. Perseus Digital Library (Tufts University) (1999.04.0007). https://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0007

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231. https://aclanthology.org/2006.amta-papers.25/

Statistics of Gender in UD_Ancient_Greek-PROIEL. (2024). https://universaldependencies.org/treebanks/grc_proiel/grc_proiel-feat-Gender.html

Stolk, J. V. (2015). Dative by Genitive Replacement in the Greek Language of the Papyri: A Diachronic Account of Case Semantics. *Journal of Greek Linguistics*, *15*(91–121). https://doi.org/10.1163/15699846-01501001

Strong's Greek: 5101. Τίς. (1890). https://biblehub.com/greek/5101.htm

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1–25. https://doi.org/10.2307/2331554

Takahashi, K., Yamamoto, K., Kuchiba, A., & Koyama, T. (2022a). Confidence interval for micro-averaged F1 and macro-averaged F1 scores. *Applied Intelligence*, *52*(5), 4961–4972. https://doi.org/10.1007/s10489-021-02635-5

Takahashi, K., Yamamoto, K., Kuchiba, A., & Koyama, T. (2022b). Confidence interval for micro-averaged F1 and macro-averaged F1 scores. *Applied Intelligence*, *52*(5), 4961–4972. https://doi.org/10.1007/s10489-021-02635-5

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1452

The PROIEL Treebank. (2025). *Proiel/proiel-treebank* [Computer software]. https://github.com/proiel/proiel-treebank (Original work published 2014)

The Second Workshop on Ancient Language Processing @NAACL 2025. (2025, August 10). https://www.aclweb.org/portal/content/second-workshop-ancient-language-processing-naacl-2025

Tolegen, G., Toleu, A., & Mussabayev, R. (2024). Contrastive Learning for Morphological Disambiguation Using Large Language Models in Low-Resource Settings. *Applied Sciences*, *14*(21), 9992. https://doi.org/10.3390/app14219992

Tsarfaty, R., Seddah, D., Kübler, S., & Nivre, J. (2013). Parsing Morphologically Rich Languages: Introduction to the Special Issue. *Computational Linguistics*, *39*(1), 15–22. https://doi.org/10.1162/COLI_a_00133

Tzanoulinou, D., Triantafyllopoulos, L., & Verykios, V. S. (2025). Harnessing Language Models for Studying the Ancient Greek Language: A Systematic Review. *Machine Learning and Knowledge Extraction*, *7*(3), Article 3. https://doi.org/10.3390/make7030071

UD Ancient Greek PROIEL. (2025). https://universaldependencies.org/treebanks/grc_proiel/index.html

UD_Ancient_Greek-PROIEL Dev Split. (2025). https://github.com/UniversalDependencies/UD_Ancient_Greek-PROIEL/blob/master/grc_proiel-ud-dev.conllu

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wu, T., Ribeiro, M. T., Heer, J., & Weld, D. (2019). Errudite: Scalable, Reproducible, and Testable Error Analysis. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 747–763). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1073

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser,

Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation* (arXiv:1609.08144). arXiv. https://doi.org/10.48550/arXiv.1609.08144

Wycliffe Financial Accountability. (2025, August 10). https://www.wycliffe.org/about/financial-accountability

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., … Li, J. (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In J. Hajič & D. Zeman (Eds.), *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 1–19). Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-3001

Appendix A

This appendix provides supplementary data and additional details related to the results as outlined in Chapter 4—Results. Specifically, it provides an example of a raw pipeline execution log as shown in Figure 18. Data related to specific features selected for each random seed for the high-risk slice identification is provided in Table 24. A pseudo code algorithm of the DART-FT methodology is included in Figure 19. Figure 20 provides a detailed depiction of the experimental pipeline. Additionally, reference data with respect to the time periods and evolution of the Ancient Greek language is provided for additional context.

```
cat overnight-20251213.yolo.log | grep "\[mine\]" | grep -v building
[mine] risky_dev_fvs=6 train_intersect=6 test_mask_sum=487 skipped_by_cap=6
max_fv_share=0.35
[mine] feature mix: {'Mood': 2, 'Case': 1, 'Gender': 1, 'Poss': 1, 'PronType': 1}
[mine] risky_dev_fvs=9 train_intersect=9 test_mask_sum=639 skipped_by_cap=6
max_fv_share=0.35
[mine] feature mix: {'Mood': 2, 'Tense': 2, 'Case': 1, 'Degree': 1, 'Gender': 1, 'Poss': 1,
'PronType': 1}
[mine] risky_dev_fvs=8 train_intersect=8 test_mask_sum=621 skipped_by_cap=6
max_fv_share=0.35
[mine] feature mix: {'Tense': 2, 'Case': 1, 'Degree': 1, 'Gender': 1, 'Mood': 1, 'Poss': 1,
'PronType': 1}
[mine] risky_dev_fvs=6 train_intersect=6 test_mask_sum=487 skipped_by_cap=8
max_fv_share=0.35
[mine] feature mix: {'Case': 1, 'Degree': 1, 'Gender': 1, 'Poss': 1, 'PronType': 1, 'Tense': 1}
[mine] risky_dev_fvs=8 train_intersect=8 test_mask_sum=636 skipped_by_cap=6
max_fv_share=0.35
```

[mine] feature mix: {'Mood': 2, 'Case': 1, 'Degree': 1, 'Gender': 1, 'Poss': 1, 'PronType': 1, 'Tense': 1}

Figure 18 Raw execution logs of the high-risk slice identification pipeline across the 5 random seeds. The output indicates which and the number of features were included in the high-risk slice, the intersection of the features between the validation and training sets, the number of features excluded from the slice due to the cap, and the total counts of tokens from the test set in the slice.

```
FOR seed s in SEEDS DO
  # (1) Baseline (CE, no oversampling)
  SetSeed(s)
  MODEL_base = InitFromPretrained()
  Train(MODEL_base, TRAIN, loss=CE)
  PRED_dev = Predict(MODEL_base, DEV)

  # (2) Pass 1 — Mine high-error FV patterns on DEV dataset
  CANDIDATES = {}
  FOR each FV seen in DEV gold labels DO
      count = freq(FV in DEV gold)
      err   = freq(FV missing from baseline prediction when FV is in gold)
      IF (err/count >= MIN_ERR) AND (count >= MIN_FREQ) THEN
          Add FV to CANDIDATES with key=err
      ENDIF
  ENDFOR

  # (3) Pass 2 — Greedy diversity cap (marginal share)
  Sort CANDIDATES by descending error count
  HIGH_RISK_FVS = {}
  SLICE_TOKENS  = {}   # DEV token indices in current slice union

  FOR FV in CANDIDATES DO
      TOK = DEV tokens whose gold contains FV
      new_union = SLICE_TOKENS ∪ TOK
      marginal_share = (|new_union| - |SLICE_TOKENS|) / |new_union|
      IF marginal_share <= MAX_SHARE THEN
          Add FV to HIGH_RISK_FVS
          SLICE_TOKENS = new_union
      ENDIF
  ENDFOR

  # (4) Pass 3 — Freeze slice + stamp masks (gold-based)
  MASK_train = (TRAIN token gold contains any FV in HIGH_RISK_FVS)
  MASK_test  = (TEST  token gold contains any FV in HIGH_RISK_FVS)

  # (5) Targeted fine-tuning grid
  FOR OS in OVERSAMPLE_FACTORS tested DO
    FOR LOSS in {CE, CB(beta), CB+F(beta,gamma)} DO

      TRAIN_os = OversampleSentences(TRAIN, MASK_train, factor=OS)
```

```
      MODEL_tgt = CopyWeights(MODEL_base)      # warm start
      Train(MODEL_tgt, TRAIN_os, loss=LOSS)

      # Evaluate (high-risk slice + global)
      EvalHighRisk(MODEL_tgt, TEST, MASK_test) # bundle-accuracy + macro-F1
      EvalGlobal(MODEL_tgt, TEST)              # global macro-F1

      # Paired test vs baseline on high-risk slice
      McNemarExact(baseline vs targeted correctness on MASK_test tokens)

    ENDFOR
  ENDFOR
ENDFOR
```

Figure 19 Pseudo code algorithm of the DART-FT methodology

Table 24 High-risk feature-value pairs for each of the 5 random seeds

Feature-Value Pair	Seed 41	Seed 42	Seed 43	Seed 44	Seed 45
Gender=Masc,Neut	X	X	X	X	X
PronType=Int	X	X	X	X	X
Case=Voc	X	X	X	X	X
Poss=Yes	X	X	X	X	X
Mood=Sub	X				X
Mood=Opt	X	X			
Mood=Imp		X	X		X
Tense=Fut		X	X	X	
Degree=Cmp		X	X	X	X
Tense=Pqp		X	X		X

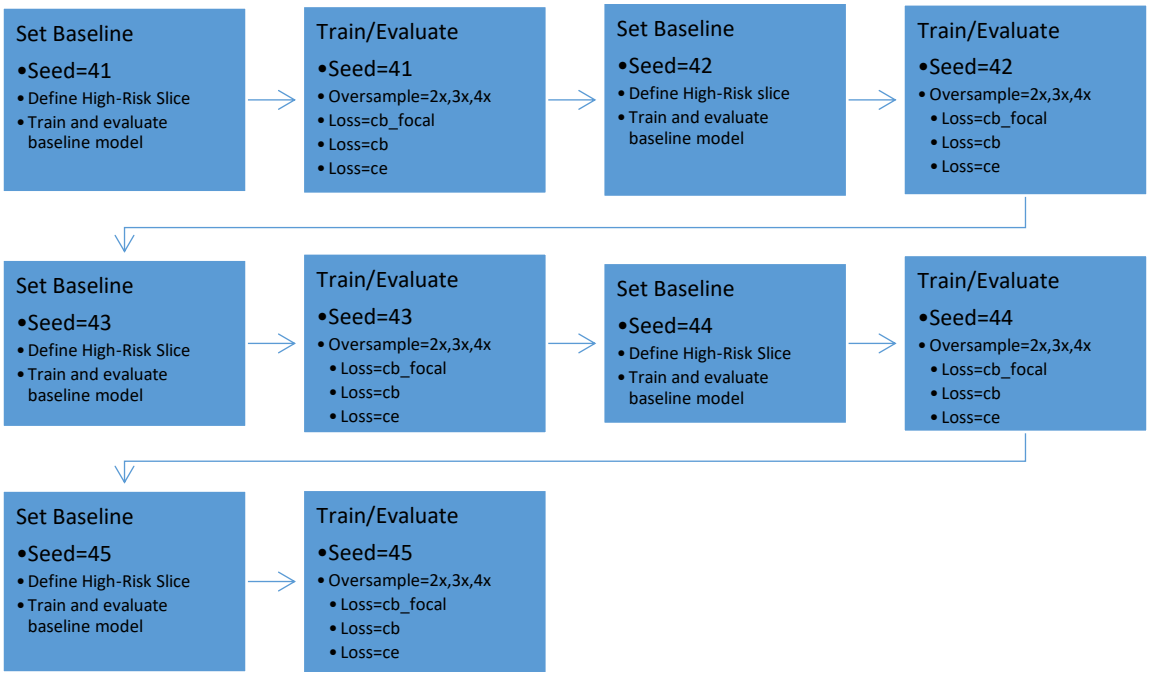

Figure 20 Experimental pipeline details which shows baseline training, high-risk slice definition, and targeted fine-tuning and evaluation for each random seed

Ancient Greek Language Time Periods and Evolution

A brief and non-exhaustive review of the history of Ancient Greek language development is useful for contextual understanding. The origins of Ancient Greek language date to before the eighth century BCE and evolved through time. Ancient Greek is categorized by scholars into several distinct time periods, some of which can overlap. These general categories are outlined in Table 25. It is also important to note that additional designations also exist, such as Koine (κοινή) or common Greek, which spans the third century BCE to the sixth century CE (Aitken & Dhont, 2023, p. 434). Koine Greek is the language in which the New Testament was penned and most of the corpus within the PROIEL treebank. In this study, Ancient Greek refers to both Koine and Classical Ancient Greek, to align with the contents of the PROIEL treebank which includes both the New Testament and texts from Herodotus.

Table 25 Time Periods of Ancient Greek Language (Bentein, 2016, p. 6).

Time Period (Centuries)	Period Name
8-6 BCE	Archaic Greek
5-4 BCE	Classical Greek
3-1 BCE	Early Post-Classical Greek
1-3 CE	Middle Post-Classical Greek
4-6[8] CE	Late Post-Classical Greek
7-8 CE	Early Byzantine Greek

The time periods contain slight differences in language. For example, between the third century BCE and the eighth century CE, the grammatical dative case was gradually replaced by the genitive case (Stolk, 2015, p. 92). These transitions in language patterns may impact model weights if the training corpus spans multiple time periods.

[8] Some scholars consider Post-Classical Greek to span from 323BCE to 1453CE (la Roi, 2020, p. 214).