Perplexity as a Metric for Dialectal Distance: A Computational Study of Greek Varieties

Stergios Chatzikyriakidis[1*], **Erofili Psaltaki**[2*],
Dimitrios Papadakis[1], **Erik Henriksson**[2], **Veronika Laippala**[2]
[1]University of Crete, [2]TurkuNLP, University of Turku
stergios.chatzikyriakidis@uoc.gr, erofili.psaltaki@utu.fi,
philp0961@philology.uoc.gr, mavela@utu.fi, erik.henriksson@utu.fi

Abstract

In this paper, we use LLM perplexity as a measure to assess Greek dialectal distance[1]. We test seven models on Standard Modern Greek (SMG) and eight dialects, namely Heptanesian, Cypriot, Maniot, Pontic, Northern, Cretan, Tsakonian, and Griko. Using samples of 5k, 15k, and 25k tokens from the GRDD+ corpus for each variety, we find a consistent dialect ranking across models, with Heptanesian closest to SMG, and Griko most distant (perplexity ratio 3.6–14.5× depending on model). These results are largely in agreement with theoretical dialectological knowledge. For example, Tsakonian consistently appears distant in all measures, reflecting its status as the sole Doric descendant, while Heptanesian appears closer by all metrics, pointing to its status as one of the dialects used to shape the official variety. Perplexity correlates strongly with Bits-Per-Character (mean r = 0.94) and Normalized Compression Distance (mean r = 0.87, range 0.76–0.93), providing support for its use as a dialectometric tool. However, a number of important confounds are also found. First, tokenization effects compress Llama 2's perplexity range. Second, genre artifacts seem to inflate the results for Cretan. Third, potential training data contamination likely reduces perplexity for Cypriot and Pontic. Lastly, we find that Greek-specific models like Meltemi and Krikri do not consistently outperform general models.

1 Introduction

Modern Greek dialects provide an interesting case study for distance studies due to their rich dialectal diversity, ranging from varieties closely related to Standard Modern Greek (SMG), such as Heptanesian, to highly divergent varieties like Tsakonian and Griko, the latter spoken in Southern Italy and heavily influenced by Romance languages. Trudgill (2003) claims that Tsakonian, Southern Italian Greek, Pontic, and Cappadocian are the varieties that are linguistically very different from Standard Greek. While most research on Greek dialects has focused on their geographical distribution and classification, there remains a gap in computationally assessing their linguistic distances and classifying them based on these distances. Even in traditional dialectology, there is no classification of Greek dialects based on their linguistic distance.

Recently, the computational assessment of linguistic distance between language varieties has gained renewed interest with the advent of large language models (LLMs). In this context, automatic metrics such as LLM perplexity can be used to approximate linguistic distances—both inter-dialect distance and distance from the standard variety—regardless of how much the model has been exposed to each variety during training. Employing these metrics contributes both to our understanding of dialect relationships and to the development of low-cost methods in computational dialectology.

This study examines Greek dialectal distances using LLM perplexity as a primary metric, complemented by information-theoretic measures (bits per character, BPC) and compression-based distances (normalized compression distance, NCD). Together, these metrics offer complementary insights into dialectal distance: LLM perplexity captures divergence from patterns learned by the model, bits per character (BPC) provides a measure of predictability and normalized compression distance (NCD) quantifies similarity based on shared redundancy.

2 Related Work

A foundational work on Modern Greek dialects is that by Newton (1972), who established that high vowel loss and vowel raising define the northern

[*] These authors contributed equally.

[1] The full code for the dialect metrics and the dataset are available at the following anonymous link: `https://github.com/TurkuNLP/Greek_dialects_perplexity`.

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 101–112
March 29, 2026.

dialects. Another foundational work is Trudgill (2003). His cartographical representation identifies fifteen dialect areas based on six phonological criteria. Kontossopoulos (2001) distinguishes *dialektoi* (Tsakonian, Pontic, Cappadocian, Southern Italian Greek) from *idiomata* (all other varieties), a taxonomy predicting that the former should show substantially greater distance from SMG. Among divergent varieties, Tsakonian is the sole Modern Greek descendant of Doric rather than Koine Greek (Nicholas, 2019), while Griko preserves the infinitive, lost elsewhere by the Medieval period (Holton and Manolessou, 2010) alongside extensive Romance contact effects (Manolessou, 2005). Pontic retains distinct features including infinitive constructions and distinctive across-the-board enclitic placement (Sitaridou, 2014; Condoravdi and Kiparsky, 2002). Cypriot exhibits phonological differences in terms of gemination (Arvaniti, 2001) plus a number of morphosyntactic differences including clefts and differences in clitic positioning.

Previous research in NLP has explored the use of smaller models for perplexity-based pruning (Ankner et al., 2024), along with analyses of dialectal biases in LLMs (Pan et al., 2025). Other studies have assessed the fairness and robustness of LLMs in handling dialects across canonical reasoning tasks (Lin et al., 2025) and have investigated tokenization and representation biases in multilingual models for dialectal NLP tasks (Kanjirangat et al., 2025). Additional work includes the creation of a library for exploring where a particular language model is perplexed (Cooper and Scholak, 2024), in addition to analyses of low-perplexity sequences—high-probability text spans generated by language models (Wuhrmann et al., 2025). Furthermore, perplexity has also been used as a stylistic signal for authorship attribution through author-specific language models (Huang et al., 2025).

There is also research on dialect metrics, such as the application of compression-based similarity measures to the quantification of distances among Bulgarian dialects (Simov and Osenova, 2007), and model training with optimal tokenization levels in datasets with consistent and inconsistent writing practices (Kanjirangat et al., 2023). Other studies have attempted to predict dialectal features at the token level for Norwegian dialects (Barnes et al., 2023).

Previous research on the classification of Modern Greek dialects has largely focused on their geographical distribution, with the most recent classification proposed by Trudgill. There have also been classification attempts for Athenian Greek and Cypriot Greek based on vowel acoustic parameters (Themistocleous, 2017), as well as on measures of temporal and spectral information from selected consonants (Themistocleous, 2019). More recent work has quantified dialectal distances among Asia Minor Greek dialects using dialectometric techniques (Bompolas and Melissaropoulou, 2025). To the best of our knowledge, this study is the first attempt to computationally measure Greek dialectal distances using multiple metrics.

3 Data

We utilize the GRDD+ dataset (Chatzikyriakidis et al., 2025), an extended Greek dialectal dataset that includes data from ten varieties of Greek. Four of these dialects—Cretan, Cypriot, Pontic, Northern— and Standard Modern Greek were part of the original GRDD dataset (Chatzikyriakidis et al., 2023) and have since been expanded in terms of coverage by the dataset's authors. Six additional varieties have been incorporated into the GRDD+: Greco-Corsican, Griko, Heptanesian, Tsakonian, Maniot, and Katharevusa. For the purposes of the present study, we use all available varieties of the GRDD+ except Greco-Corsican, as it is extinct, and Katharevusa, since it represents a historical formal register.

Standard Modern Greek (SMG) is the official language of both Greece and Cyprus (Mackridge, 1985). It is based primarily on Peloponnesian dialects and was later enriched by features from the dialects of Istanbul and the Ionian Islands.

Cypriot Greek is the native language of most Greek Cypriots, both on the island of Cyprus and in the diaspora, although the official language of the Republic of Cyprus is Standard Modern Greek (SMG) (Newton, 2013; Tsiplakou, 2014).

Pontic Greek is a Modern Greek dialect that ultimately derives from Koine Greek and is historically associated with the Pontus region on the southern Black Sea coast (Sitaridou and Chatzikyriakidis, 2012; Schreiber, 2018).

Cretan Greek is the variety spoken on the island of Crete. Like many other Modern Greek dialects, it descends from Koine Greek (Mackridge, 1985).

Northern Greek is the form of Greek spoken in Thessaly, Thrace, and the Northeast Aegean.

Griko is a Greek variety spoken in Grecìa Salentina in southern Italy, where it is recognized as a minority language (Salminen, 1999). The dialect is written with the Latin alphabet (Chatzikyriakidis, 2010).

Heptanesian is a Modern Greek dialect spoken in the Ionian Islands. The dialect shows strong Venetian and Italian influence due to centuries of Venetian rule (Ralli, 2012).

Tsakonian, still spoken in the eastern Peloponnese, is a highly divergent Modern Greek variety considered by many to be a separate language, exceptional for descending directly from the ancient Doric dialect (Joseph et al., 1987; Mackridge, 2010).

Maniot is the variety spoken in the region of Laconian Mani in the southern Peloponnese (Trudgill, 2003).

4 Models Evaluated

We evaluate seven open-weight model variants, primarily drawn from the Llama family together with two Greek-specialized checkpoints. Open-weight models are required because perplexity computation relies on access to token-level log probabilities, which closed API models (closed GPT models, Claude, Gemini) do not expose. Although several systems share the same underlying architecture, they differ substantially along dimensions that are directly relevant to dialect modeling: training generation, parameter scale, and language specialization.

All models are evaluated strictly in a zero-shot setting. We use the publicly available checkpoints as released, without additional fine-tuning, continued pretraining, or exposure to Standard Modern Greek or dialectal evaluation data. Consequently, perplexity differences reflect the models' inherent pretrained linguistic coverage rather than task-specific adaptation.

Our selection spans three dimensions. First, model generation: Llama 2 represents an earlier architecture with limited Greek tokenization, while Llama 3.x models show improved multilingual coverage. Second, model scale: we compare 7–8B parameter models with 70B variants to assess whether scale improves dialectal discrimination. Third, language specialization: Krikri-8B and Meltemi 7B are explicitly trained on Greek data, allowing us to test whether Greek-focused pretraining improves dialectal coverage.

- **General-purpose:** Llama 2 7B, Llama 3 8B, Llama 3.1 8B, Llama 3.1 70B, Llama 3.3 70B
- **Greek-specialized:** Krikri-8B[2], Meltemi 7B[3]

5 Metrics and Diagnostics

We incorporate multiple complementary metrics to capture different aspects of linguistic complexity and structural similarity. These measures fall into two groups: metrics that rely on LLM outputs and metrics computed directly from the raw text and tokenization, without any LLM inference. The only LLM-based metric is perplexity, computed across seven open-weight models using token-level log probabilities. The remaining metrics, Bits-Per-Character (BPC), Normalized Compression Distance (NCD), and Tokens per Word, are calculated solely from the input strings and their tokenization. These metrics capture different aspects of linguistic complexity and structural similarity. Each metric is described below.

- **Perplexity (PPL).** Measures how "surprised" a language model is by the text, defined as the exponential of cross-entropy loss. Lower PPL means the text is closer to the model's training distribution. Perplexity is not an intrinsic property of a dialect, rather, it measures distance from what the model has learned as "Greek". It is therefore model-dependent: a model trained on Pontic would show Pontic as closer and SMG as more distant. PPL is also confounded by genre, register and topic.

- **Bits-Per-Character (BPC).** An information-theoretic normalization of perplexity, defined as: $\text{BPC} = \log_2(\text{PPL}) \times \frac{\text{tokens}}{\text{characters}}$. BPC accounts for tokenization differences across models. A model with poor Greek tokenization (i.e., many tokens per word) would otherwise exhibit artificially high perplexity. BPC enables cross-model comparison. It shares the same theoretical limitations as PPL but is more robust to tokenizer effects.

- **Normalized Compression Distance (NCD).** Approximates Kolmogorov complexity by measuring shared information between texts using compression algorithms. NCD is model-agnostic and does not depend on LLM training data. While theoretically well-grounded, the NCD values in our dataset are tightly clustered (range 0.988–1.020), making it difficult

[2] Built on Llama 3.1 8B and further trained on Greek data.

[3] Developed by the same team as Krikri; they recommend Krikri 8B as their newest model.

to distinguish closely related dialects. It performs better for more distant languages, and is therefore included primarily for completeness, as PPL and BPC are more discriminative in this setting.

- **Tokens per Word.** A diagnostic metric measuring tokenizer coverage. Lower values indicate better vocabulary match. For SMG, values are 1.46 for Greek-specialized models (Krikri/Meltemi) versus 6.12 for Llama 2. The near-6 tokens-per-word score of Llama 2 reflects character-level fallback, indicating that the tokenizer fails to recognize Greek words.

- **Characters per Token.** The inverse of fragmentation, with higher values indicating more efficient subword units. SMG shows 4.4 characters per token for Greek-specialized models versus 1.05 for Llama 2. Values near 1.0 signal byte-level tokenization. Greek-specialized tokenizers (Krikri, Meltemi) achieve over four characters per token for SMG, dropping to approximately 2.3 for Griko, showing recognition of SMG morphemes but fragmentation of dialectal vocabulary.

- **Unique Token Ratio.** The proportion of unique tokens in the corpus serves as an indicator of lexical diversity, with higher values reflecting more diverse vocabulary usage. For SMG, this proportion is 21% for Krikri, compared to only 0.5% for Llama 2. Such low values indicate that the model repeatedly uses a very limited set of tokens, which is a typical signature of character-level fallback, where tokenization relies primarily on individual characters rather than lexical units. This measure therefore functions as a diagnostic for tokenizer quality.

- **Coefficient of Variation (CV) of PPL.** Defined as the standard deviation divided by the mean of window-level perplexity, measuring text homogeneity. Low CV indicates consistent text (e.g., Cretan rhymed folkloric material: 0.22–0.24), while high CV reflects a heterogeneous corpus (e.g., Pontic: 0.81–0.90). The low CV for Cretan confirms genre homogeneity, meaning that all windows are similarly surprising because they belong to the same poetic style.

- **Inter-metric Correlations.** We examined the relationships between our different metrics to assess whether they provide consistent signals about dialect similarity. The high correlation between PPL and BPC (mean r = 0.94 across models) is partly definitional, as BPC is derived from PPL. The correlation between PPL and NCD (mean r = 0.87, range 0.76–0.93) is more informative, indicating convergence between independent methods. The agreement of multiple metrics on the same dialect ordering (Heptanesian closest, Griko and Tsakonian most distant) provides stronger evidence than any single metric alone.

  PPL–NCD correlations were computed between model perplexities and a model-agnostic Normalized Compression Distance vector calculated relative to Standard Modern Greek (NCD(dialect, SMG)); correlations are r unless otherwise noted.

6 Results

6.1 Perplexity Across Models

Table 1 presents perplexity measurements across all seven models. As expected, Standard Modern Greek obtains the lowest perplexity for all models, reflecting its dominance in training corpora. Perplexity increases systematically for dialects that diverge more substantially from SMG, with the highest values for Tsakonian and Griko, both highly divergent varieties. The dialect ranking is quite consistent: Heptanesian appears closest to SMG, followed by Cypriot and Maniot, then Pontic and Northern, with Cretan, Tsakonian, and Griko showing the highest distances.

Model-wise, even the largest models (Llama 3.1 70B and Llama 3.3 70B) show substantial gaps between SMG and the most divergent dialects. The Greek-specialized models (Krikri and Meltemi) offer small gains for some dialects but do not consistently outperform general-purpose models and still yield high perplexity overall. This suggests that Greek-focused training data does not ensure broad dialectal coverage, likely due to limited dialectal data. An anomaly appears in Meltemi, where Heptanesian scores unexpectedly lower than SMG. This finding may be due to the close similarity to SMG, given that the model has not been trained on dialectal data.

The Griko/SMG ratio ranges from 3.6× (Llama 2 7B) to 14.5× (Llama 3.1 8B), reflecting differences in tokenization and Greek variety coverage, and underscoring the extreme difficulty these highly distinct dialects pose for the models.

6.2 Tokenization Effects

Table 2 shows clear differences in tokenization efficiency across both dialects and models. SMG consistently receives the fewest tokens per word, while more distinct varieties like Northern, Pontic, and Griko are segmented less efficiently. General-purpose models (Llama variations) show limited sensitivity to Greek morphological structure, whereas Greek-specialized tokenizers (Krikri, Meltemi) produce both lower overall token counts and a distribution that aligns with the expected dialectal distances.

The patterns help contextualize the perplexity results. With 6.12 tokens per word for SMG, compared to 1.46 for Greek-specialized models, Llama 2 operates at near-character level. The model predicts characters rather than linguistic units, compressing the perplexity range and reducing sensitivity to dialectal structure. However, even in this way, the perplexity measures here follow the same pattern as the rest of the models in terms of dialectal distance ranking.

Greek-specialized tokenizers (Krikri, Meltemi) show efficient tokenization for SMG (1.46–1.47 tok/word) that degrades for more distant dialects (2.40–2.44 for Griko). This 1.67× decrease in tokenization efficiency contributes directly to the larger perplexity gaps observed in these models, reflecting their greater sensitivity to genuine linguistic divergence. Although tokens per word provide insight into tokenization efficiency, we did not compute a quantitative baseline such as token overlap between dialects. Incorporating such a measure in future work could more precisely relate tokenization patterns to genuine linguistic divergence.

6.3 Model Comparison

Table 3 summarizes model performance and presents a comparison between the variety with the lowest perplexity (SMG) and the highest perplexity (Griko), based on a sample of 25,000 words. The results show a consistent and substantial performance gap between the two varieties for every model.

Across the board, SMG yields consistently lower perplexity values, reflecting its status as the standard variety and its strong representation in training data. Griko, on the other hand, systematically exhibits the highest perplexity, confirming its position as the most challenging dialect for all models. Despite the transliteration of its Latin alphabet into Greek, this heavily Italian-influenced variety differs significantly from the other varieties.

The PPL ratio quantifies this contrast, revealing that modeling Griko is between 3.6× and 14.5× more difficult than modeling SMG, depending on the architecture. Llama 3.1 8B shows the highest sensitivity (14.5× ratio, 2.47 BPC range), while Llama 2 7B shows the lowest (3.6×, 1.83). The 70B models show slightly lower ratios than their 8B counterparts, a fact that might point to memorization effects that smooth over dialectal irregularities. In addition, there is a substantial gap in overall perplexity between the smallest Llama model and the larger variants, with the 7B model consistently yielding higher scores across conditions. This difference is likely attributable to model scale, as smaller models have more limited representational capacity, resulting in weaker probability estimates overall. The pattern aligns with well-established scaling effects in language modeling, where increases in parameter count systematically improve perplexity.

The Greek-specific models (Krikri-8B, Meltemi 7B) still display large performance gaps between SMG and Griko, since they exhibit very high perplexity on Griko and large PPL ratios, indicating limited robustness to this minority variety. While Meltemi 7B shows a comparatively lower ratio, this is probably driven by its unusually high perplexity on SMG rather than improved modeling of Griko. Overall, these results suggest that specialization in Greek does not automatically translate into effective coverage of all the varieties, confirming the underrepresentation of such dialects even in language-specific models.

Bits-Per-Character (BPC) normalizes perplexity for tokenization differences, enabling fairer cross-model comparison. The BPC range (Griko minus SMG) is relatively consistent across models (1.83–2.47) compared to raw perplexity ratios, indicating that part of the observed variation indeed stems from differences in tokenization. However, the stability of the BPC range across architectures also suggests that the persistent performance gap between SMG and Griko points to deeper lexical, morphological, and distributional mismatches between the two varieties.

Dialect	General-Purpose Models					Greek-Specialized	
	Llama 2 7B	Llama 3 8B	Llama 3.1 8B	Llama 3.1 70B	Llama 3.3 70B	Krikri 8B	Meltemi 7B
SMG	2.91	5.17	6.15	4.69	4.81	8.83	39.43
Heptanesian	3.76	8.87	11.74	8.41	8.83	12.36	25.38*
Cypriot	4.79	18.18	25.40	17.50	18.48	24.74	51.86
Maniot	5.12	18.66	25.91	18.62	20.07	24.73	55.59
Pontic	6.47	23.09	32.05	23.57	24.55	37.15	48.95
Northern	7.53	33.32	49.62	36.13	37.47	41.63	57.67
Cretan†	6.72	41.35	63.02	45.96	48.08	62.07	123.92
Tsakonian	7.98	46.81	67.61	51.41	52.75	96.54	192.55
Griko‡	10.46	59.66	89.32	59.45	61.86	121.57	208.76
Ratio (G/S)	**3.6×**	**11.5×**	**14.5×**	**12.7×**	**12.9×**	**13.8×**	**5.3×**

Table 1: Perplexity (PPL) by Model and Dialect (25,000 words).

†Primarily rhymed folkloric material. ‡Transliterated from Latin. *Lower than SMG (anomaly). Ratio = Griko/SMG.

Dialect	Llama 2	Llama 3.x	Krikri	Meltemi
SMG	6.12	2.38	1.46	1.47
Heptanesian	5.79	2.41	1.63	1.63
Cypriot	6.04	2.62	1.85	1.86
Maniot	5.78	2.60	1.94	1.91
Cretan	5.33	2.40	1.83	1.79
Pontic	5.44	2.84	2.17	2.24
Northern	5.86	2.99	2.38	2.45
Tsakonian	5.39	2.73	2.12	2.11
Griko	5.51	3.00	2.44	2.40
Ratio (G/S)	**0.90×**	**1.26×**	**1.67×**	**1.63×**

Table 2: Tokenization: Tokens per Word (25,000 words).

Llama 3.x = shared tokenizer across Llama 3, 3.1, 3.3.

6.4 Inter-Metric Correlations

Table 4 shows correlations between PPL, BPC, and NCD across all seven models. PPL and BPC are strongly correlated (mean $r = 0.94$), which is partly definitional since BPC is derived from PPL. The PPL–NCD correlation (mean $r = 0.87$, range 0.76–0.93) is more informative, as it links two independent methods: neural language modeling and compression-based distance. This convergence suggests that both metrics capture similar distributional properties. Greek-specialized models (Krikri, Meltemi) show somewhat lower PPL–NCD correlations, possibly reflecting their different tokenization strategies.

6.5 Stability Across Sample Sizes

We evaluated all models at 5,000, 15,000, and 25,000 words. Table 5 presents the coefficient of variation (CV) across sample sizes for all models. The dialect ranking remains stable across sample sizes, with most CV values below 10%.

The most distant varieties show the highest stability: Griko (CV 0.1–2.7%) and Tsakonian (0.2–3.2%) yield consistent estimates across all models. Cretan also shows low variance (1.0–3.9%), reflecting the stylistic homogeneity of a corpus, a substantial part of which consists

Model	SMG PPL	Griko PPL	PPL Ratio	BPC Range
Llama 2 7B	2.91	10.46	3.6×	1.83
Llama 3 8B	5.17	59.66	11.5×	2.25
Llama 3.1 8B	6.15	89.32	14.5×	2.47
Llama 3.1 70B	4.69	59.45	12.7×	2.30
Llama 3.3 70B	4.81	61.86	12.9×	2.32
Krikri-8B	8.83	121.57	13.8×	2.28
Meltemi 7B	39.43	208.76	5.3×	2.06

Table 3: Model Comparison Summary (25,000 words).

BPC Range = Griko BPC − SMG BPC.

Model	PPL–BPC	PPL–NCD	BPC–NCD
Llama 2 7B	.99***	.93***	.96***
Llama 3 8B	.95***	.90**	.96***
Llama 3.1 8B	.94***	.90**	.96***
Llama 3.1 70B	.95***	.90**	.96***
Llama 3.3 70B	.94***	.90**	.96***
Krikri-8B	.93***	.83*	.94***
Meltemi 7B	.89**	.76*	.93***
Mean	**.94**	**.87**	**.95**

Table 4: Inter-Metric Correlations by Model ($n = 8$ dialects).

*$p < .05$, **$p < .01$, ***$p < .001$.

mainly of rhymed folkloric material. Northern shows the highest variance across most models (CV 9.7–12.4%), pointing to a more heterogeneous corpus. Similarly, Pontic shows elevated variance for Greek-specialized models (8.2–12.3%), likely reflecting the diversity of sources: Wikipedia articles, theatrical plays, jokes, and songs.

Meltemi shows anomalously high CV for SMG (22.9%), consistent with the corpus composition issues noted earlier. General-purpose models show lower variance for SMG (2.2–4.2%) than Greek-specialized models, possibly reflecting more stable SMG representation in their training data.

Table 6 confirms that the Griko/SMG ratio is stable across sample sizes for most models. Llama 2 7B shows consistent low sensitivity (3.3–3.6×), while Llama 3.x models cluster around 11–14×. The notable exception is Meltemi, whose ratio drops from 9.4× at 5k to 5.3× at 25k. The 25,000-word results represent the most reliable estimates, as increasing sample size reduces sampling variability and leads to the stabilization of frequency-based measures.

7 Discussion

7.1 Dialect-Specific Findings

Tsakonian is shown to be the second most distant variety in our findings (PPL 47–193 across models). This is in line with its status in Greek dialectology. For example, this aligns with Nicholas's (2019) lexicostatistical analysis that dates the divergence of Tsakonian back to approximately 800 CE and, furthermore, seems to reflect fundamental structural differences stemming from this divergence. Notably, Tsakonian shows better tokenization efficiency than Northern or Griko, indicating that its extreme perplexity reflects true divergence rather than tokenization artifacts.

Griko shows the highest distance across all models (PPL 59–209), reflecting centuries of complete geographic isolation following Byzantine decline in Southern Italy combined with extensive Romance contact (Manolessou, 2005). Transliteration from Latin script may introduce additional artifacts, though the consistency of Griko's extreme distance across models suggests this effect is secondary. Additional experiments using the original Latin script texts yielded almost identical results, indicating that script choice does not substantially affect the measurements.

Cretan shows anomalously high perplexity (PPL 41–124) despite efficient tokenization close to SMG levels. We attribute this to corpus composition rather than linguistic distance: the Cretan data consists primarily of rhymed folkloric material, such as mantinades, traditional rhyming couplets with fixed meter an d formulaic expressions. The models are surprised by poetic structure, not dialectal features. Supporting this interpretation, Cretan shows the lowest coefficient of variation across all models (0.22–0.24), indicating homogeneous text that is consistently surprising.

Cypriot appears surprisingly close to SMG (PPL 18–52) despite preserving substantial structural distinctiveness, including phonemic geminates (Arvaniti, 2001) and morphosyntactic conservatism (Newton, 1972). These features may be underrepresented in orthographically normalized text, and Cypriot's substantial online presence as the native variety of Cyprus, likely reduces model surprise through training exposure. This contamination hypothesis requires empirical verification.

Heptanesian is closest to SMG across all models (PPL 8–25), confirming its historical contribution to standardization. Trudgill (2003) notes that the Ionian Islands supplied most of the input into Standard Greek, and the Heptanesian School of literature directly shaped the emerging Demotic standard. The Meltemi result (Heptanesian < SMG) potentially reflects corpus composition issues rather than genuine linguistic relationships, but needs further investigation.

Pontic occupies a mid-range position despite traditional classification as highly divergent. The variety preserves infinitive constructions (Sitaridou, 2014), SOV vestiges, distinctive clitic placement (Condoravdi and Kiparsky, 2002), and archaic negation (Drettas, 1997), yet shows only moderate perplexity. Training data exposure offers a plausible explanation: Pontic has its own Wikipedia likely included in LLM training, and its poor tokenization efficiency (which typically inflates perplexity) fails to push scores higher, consistent with contamination reducing apparent distance.

Northern varieties show mid-range distance consistent with their classification as *idiomata* rather than *dialektoi* (Kontossopoulos, 2001). The defining vowel phenomena (Newton, 1972; Topintzi and Baltazani, 2012) represent phonological rather than morphosyntactic divergence, and in orthographically normalized text these differences may be partially obscured.

	General-Purpose Models					Greek-Specialized	
Dialect	Llama 2 7B	Llama 3 8B	Llama 3.1 8B	Llama 3.1 70B	Llama 3.3 70B	Krikri 8B	Meltemi 7B
SMG	3.9	2.6	4.2	2.2	2.2	8.9	22.9
Heptanesian	5.0	7.9	8.8	7.5	7.7	3.6	4.2
Cypriot	4.0	5.5	6.2	4.3	4.7	1.8	5.1
Maniot	3.9	6.2	5.4	8.0	7.7	8.6	5.5
Pontic	0.8	5.1	5.0	5.6	5.4	8.2	12.3
Northern	4.9	9.7	11.4	10.7	10.8	11.0	12.4
Cretan†	1.0	2.5	2.4	3.9	3.9	1.9	1.0
Tsakonian	0.2	1.9	2.0	3.2	2.9	3.2	2.2
Griko‡	0.1	1.2	0.6	2.5	1.9	2.5	2.7

Table 5: Coefficient of Variation (%) Across Sample Sizes (5k, 15k, 25k words).

†Rhymed folkloric material. ‡Transliterated. Lower CV = more stable estimates.

Model	5k	15k	25k
Llama 2 7B	3.3×	3.6×	3.6×
Llama 3 8B	10.9×	10.7×	11.5×
Llama 3.1 8B	13.1×	13.5×	14.5×
Llama 3.1 70B	13.3×	12.7×	12.7×
Llama 3.3 70B	13.4×	12.7×	12.9×
Krikri-8B	16.2×	12.3×	13.8×
Meltemi 7B	9.4×	5.5×	5.3×

Table 6: Griko/SMG Perplexity Ratio Across Sample Sizes.

Meltemi variance reflects SMG corpus issues.

Maniot consistently occupies an intermediate position, sharing the archaic $\upsilon \rightarrow$ /u/ with Tsakonian but deriving from Koine Greek (Kontossopoulos, 2001).

7.2 Methodological Implications

Three findings have implications for perplexity as a dialectometric tool. First, tokenization effects dominate cross-model comparison: Llama 2's compressed range (3.6× ratio vs. 11–14× for others) reflects character-level tokenization rather than reduced dialectal sensitivity, and bits-per-character normalization following Mielke et al. (2018) enables fairer comparison. Second, genre confounds the dialect signal; as the Cretan case demonstrates, corpus composition can inflate perplexity independent of linguistic distance, and future work should control for this through stratified sampling. Third, training data exposure reduces perplexity in ways that may mask linguistic distance; Cypriot's and Pontic's moderate scores despite substantial traditional divergence may reflect model familiarity, and contamination detection methods (Shi et al.) could help disentangle these effects.

Our results produced rankings that are to a large degree in agreement with traditional dialectological knowledge and classification. Tsakonian's large distance can be seen as evidence of earlier branch divergence (Nicholas, 2019), while for Griko, it points to Romance contact as well as isolation (Manolessou, 2005). The minimal distance we find in Heptanesian confirms its contribution to language standardization (Trudgill, 2003). However, Pontic's mid-range position somewhat contradicts traditional classification, possibly reflecting training contamination. Cretan also shows surprisingly high perplexity. However, this likely reflects genre rather than dialect.

Lastly, the interesting finding that Greek-specialized models do not consistently outperform more general-purpose models suggests that these models have minimal dialectal knowledge, and that Greek-focused pre-training does not guarantee better results on dialectal varieties. Furthermore, this is commensurate with the observation by Fleisig et al. (2024) that LLMs exhibit standard language ideology, and thus produce degraded performance on non-standard varieties.

7.3 Validation Against Literature

Our results seem to align largely, but not fully, with the knowledge we have about these dialects from a theoretical standpoint. Heptanesian Greek is indeed considered to have proximity to SMG, given its historical contribution to the standard language. Tsakonian's distance comes as no surprise and confirms its uniqueness as the only dialect with direct descendance from Doric. Likewise, Griko's distance reflects its heavy Romance borrowing and contact. The convergence of multiple metrics (PPL, BPC, NCD) on rankings consistent with expert linguistic assessment validates perplexity as a tool for dialectometry, with appropriate caveats about genre and training data exposure.

Limitations and Future Work

The Cretan corpus consists primarily of rhymed folkloric material, such as the rhymed couplets called mantinades, introducing genre confounds that inflate perplexity independent of linguistic distance. The Griko corpus required transliteration from Latin script, potentially affecting results. Corpus sizes are limited to 25,000 words maximum in order to be more representative, given that for some dialects we do not have considerably more data than this number. In future work, we can test larger corpus sizes, at least for the dialects that do have available data. Additionally, perplexity measurements may be confounded by training data exposure: well-documented varieties like Cypriot may appear artificially close to SMG. Finally, Fleisig et al. (2024) document that LLMs show systematic bias against non-standard varieties, and thus our rankings may conflate true dialectal divergence with standard language ideology.

In future work, we aim to test the contamination hypothesis by training fine-grained Dialect Identification classifiers and examine data contamination issues by sampling data that some of our used models have been trained on, looking for dialectal contamination. Furthermore, we plan to extend the size of the datasets where possible and also create genre-controlled protocols. To complement perplexity-based metrics, future work could also explore combining LLMs with other NLP approaches, such as word embeddings or machine learning classifies, to provide more robust measures of dialectal similarity.

Acknowledgments

Stergios Chatzikyriakidis gratefully acknowledges funding from Amazon (project: Neural-Symbolic Integration for Enhanced Natural Language Processing (NIELS)) that provided computational support for the fine-tuning experiments described in the paper. Stergios Chatzikyriakidis is also partially funded by the European Union (ERC ADG, PhylProGramm, 101096554). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Erofili Psaltaki received funding from the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No 101177564—HAIF. Cofunded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Veronika Laippala and Erik Henriksson received funding from the Research Council of Finland through "FIN-CLARIAH research infrastructure" (project 358720, which has also received funding from the European Union – NextGenerationEU instrument) and "Mechanisms of register variation in massively multilingual web-scale corpora" (project 362459); Erik Henriksson additionally received funding through "Green NLP - controlling the carbon footprint in sustainable language technology" (project 353167). Furthermore, We also wish to acknowledge CSC – IT Center for Science Ltd. for providing computational resources.

References

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L Leavitt, and Mansheej Paul. 2024. Perplexed by perplexity: Perplexity-based data pruning with small reference models. *arXiv preprint arXiv:2405.20541*.

Amalia Arvaniti. 2001. Cypriot Greek and the phonetics and phonology of geminates. *MIT Working Papers in Linguistics*, 41:19–36.

Jeremy Barnes, Samia Touileb, Petter Mæhlum, and Pierre Lison. 2023. Identifying token-level dialectal features in social media. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 146–158.

Stavros Bompolas and Dimitra Melissaropoulou. 2025. Understanding dialectal variation in contact scenarios through dialectometry: Insights from inner asia minor greek. *Languages*, 10(1):13.

Stergios Chatzikyriakidis. 2010. *Clitics in four dialects of Modern Greek: A dynamic account*. Ph.D. thesis, University of London.

Stergios Chatzikyriakidis, Dimitris Papadakis, Sevasti-Ioanna Papaioannou, and Erofili Psaltaki. 2025. Grdd+: An extended greek dialectal dataset with cross-architecture fine-tuning evaluation. *arXiv preprint arXiv:2511.03772*.

Stergios Chatzikyriakidis, Chatrine Qwaider, Ilias Kolokousis, Christina Koula, Dimitris Papadakis, and Efthymia Sakellariou. 2023. Grdd: A dataset for greek dialectal nlp. *arXiv preprint arXiv:2308.00802*.

Cleo Condoravdi and Paul Kiparsky. 2002. Clitics and clause structure. *Journal of Greek Linguistics*, 2:1–39.

Nathan Cooper and Torsten Scholak. 2024. Perplexed: Understanding when large language models are confused. *arXiv preprint arXiv:2404.06634*.

Georges Drettas. 1997. *Aspects pontiques*. Association de recherches pluridisciplinaires, Paris.

Eve Fleisig, Suchin Gururangan, and Noah A. Smith. 2024. Standard language ideology in NLP. *arXiv preprint arXiv:2404.11968*.

David Holton and Io Manolessou. 2010. Medieval and early modern Greek. *Cambridge Encyclopedia of the Language Sciences*, pages 481–484.

Weihang Huang, Akira Murakami, and Jack Grieve. 2025. Attributing authorship via the perplexity of authorial language models. *PloS one*, 20(7):e0327081.

Brian D Joseph, Irene Philippaki-Warburton, and Irene Philippaki-Warburton. 1987. *Modern Greek*. Croom Helm London.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2023. Optimizing the size of subword vocabularies in dialect classification. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 14–30.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2025. Tokenization and representation biases in multilingual models on dialectal nlp tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24003–24021.

Nikolaos Kontossopoulos. 2001. *Dialektoi kai Idiomata tis Neas Ellinikis*. Gregoris, Athens. In Greek.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J Wooldridge, Janet Pierrehumbert, and Furu Wei. 2025. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342.

Peter Mackridge. 1985. *The Modern Greek Language: A Descriptive Analysis of Standard Modern Greek*. Oxford University Press, Oxford.

Peter Mackridge. 2010. Modern greek. *A Companion to the Ancient Greek Language*, pages 564–587.

Io Manolessou. 2005. The Greek dialects of Southern Italy: An overview. In *Cambridge Papers in Modern Greek*, volume 13, pages 103–125.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2018. Spell once, summon anywhere: A two-level open-vocabulary language model. In *Proceedings of NAACL-HLT*, pages 1661–1671.

Brian Newton. 1972. *The Generative Interpretation of Dialect: A Study of Modern Greek Phonology*. Cambridge University Press.

Brian Newton. 2013. *Cypriot Greek: Its phonology and inflections*, volume 121. Walter de Gruyter.

Nick Nicholas. 2019. A critical lexicostatistical examination of Ancient and Modern Greek and Tsakonian. *Journal of Applied Linguistics and Lexicography*, 1(1):18–68.

Eileen Pan, Anna Seo Gyeong Choi, Maartje Ter Hoeve, Skyler Seto, and Allison Koenecke. 2025. Analyzing dialectical biases in llms for knowledge and reasoning benchmarks. *arXiv preprint arXiv:2510.00962*.

Angela Ralli. 2012. Verbal loanblends in griko and heptanesian: a case study of contact morphology. *L'Italia Dialettale*, 73:111–132.

Tapani Antero Salminen. 1999. *UNESCO red book on endangered languages: Europe*. Helsingin Yliopisto [Host].

Laurentia Schreiber. 2018. 6.4. romeyka. In *The Languages and Linguistics of Western Asia*, pages 892–934. De Gruyter Mouton.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer.

Kiril Simov and Petya Osenova. 2007. Applying a normalized compression metric to the measurement of dialect distance. *Serdica Journal of Computing*, 1(1):73–86.

Ioanna Sitaridou. 2014. The Romeyka infinitive: Continuity, contact and change in the Hellenic varieties of Pontus. *Diachronica*, 31(1):31–73.

Ioanna Sitaridou and Stergios Chatzikyriakidis. 2012. Cultural survival shifts focus: The case of pontic greek. *When empires clash: Modern-day outcomes of historical Greek and Turkish language encounters", MedWorlds*, 4:29.

Charalambos Themistocleous. 2017. Dialect classification using vowel acoustic parameters. *Speech Communication*, 92:13–22.

Charalambos Themistocleous. 2019. Dialect classification from a single sonorant sound using deep neural networks. *Frontiers in Communication*, 4:64.

Nina Topintzi and Mary Baltazani. 2012. The acoustics of high-vowel loss in a Northern Greek dialect and typological implications. In *Consonant Clusters and Structural Complexity*, pages 369–398. De Gruyter Mouton, Berlin.

Peter Trudgill. 2003. Modern greek dialects: A preliminary classification. *Journal of Greek linguistics*, 4(1):45–63.

Stavroula Tsiplakou. 2014. How mixed is a ‘mixed’system?: The case of the cypriot greek koiné. *Linguistic Variation*, 14(1):161–178.

Arthur Wuhrmann, Andrei Kucharavy, and Anastasiia Kucherenko. 2025. Low-perplexity llm-generated sequences and where to find them. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 774–783.

A Results at Different Sample Sizes

Tables 7 and 8 present full perplexity results at 5,000 and 15,000 word sample sizes.

	General-Purpose Models					Greek-Specialized	
Dialect	**Llama 2 7B**	**Llama 3 8B**	**Llama 3.1 8B**	**Llama 3.1 70B**	**Llama 3.3 70B**	**Krikri 8B**	**Meltemi 7B**
SMG	3.15	5.49	6.82	4.71	4.83	7.59	22.16
Heptanesian	4.07	9.67	12.93	9.15	9.64	12.65	26.53
Cypriot	5.25	20.59	29.23	19.26	20.56	23.99	46.39
Maniot	4.91	17.79	25.17	17.33	18.71	24.10	54.90
Pontic	6.35	25.63	35.12	26.42	27.41	43.07	65.41
Northern	8.44	40.36	62.31	44.84	46.85	50.20	67.49
Cretan†	6.88	43.98	66.78	50.63	52.84	65.04	123.72
Tsakonian	8.02	44.73	64.53	47.60	49.26	89.16	182.70
Griko‡	10.48	59.77	89.61	62.63	64.60	123.16	208.61
Ratio (G/S)	**3.3×**	**10.9×**	**13.1×**	**13.3×**	**13.4×**	**16.2×**	**9.4×**

Table 7: Perplexity (PPL) by Model and Dialect (5,000 words).

†Primarily rhymed folkloric material. ‡Transliterated from Latin.

	General-Purpose Models					Greek-Specialized	
Dialect	**Llama 2 7B**	**Llama 3 8B**	**Llama 3.1 8B**	**Llama 3.1 70B**	**Llama 3.3 70B**	**Krikri 8B**	**Meltemi 7B**
SMG	2.90	5.44	6.56	4.92	5.05	9.43	35.99
Heptanesian	3.61	7.97	10.42	7.60	7.97	11.62	23.92
Cypriot	4.88	18.58	26.13	17.73	18.85	23.72	51.67
Maniot	5.40	20.61	28.50	21.01	22.49	29.10	61.97
Pontic	6.44	23.00	31.29	23.45	24.40	35.72	53.92
Northern	8.25	41.91	65.14	46.49	48.19	54.51	78.23
Cretan†	6.85	42.98	65.81	48.27	50.37	63.23	126.52
Tsakonian	8.01	45.35	65.29	49.09	50.33	93.00	185.65
Griko‡	10.46	58.24	88.43	62.70	64.23	116.06	197.16
Ratio (G/S)	**3.6×**	**10.7×**	**13.5×**	**12.7×**	**12.7×**	**12.3×**	**5.5×**

Table 8: Perplexity (PPL) by Model and Dialect (15,000 words).

†Primarily rhymed folkloric material. ‡Transliterated from Latin.