Teaching Logion How Scribes Actually Slip:

A Weighted Edit Distance for Byzantine Greek Emendation

Reza Ramji

Extends Logion (Cowen-Breen, Brooks, Graziosi, Haubold; *ACL ALP-2023*)

`github.com/Zed-Rez/logion-wlev`

Abstract

Princeton's Logion is a BERT-based model that flags likely scribal errors in premodern Greek and proposes corrections. To decide which spellings count as plausible alternatives to a transmitted word, it uses classical Levenshtein distance, which treats every single-character substitution as equally costly. That assumption is poorly matched to Byzantine Greek, where a small, well-documented set of phonetic and visual confusions accounts for most copying errors. We replace Logion's binary distance-1 filter with a *weighted* Levenshtein neighbourhood whose substitution costs come from the standard palaeographic and phonological literature. On real attested manuscript variants the change yields a statistically significant gain: top-5 recovery of 42.3% versus a baseline of 35.7% on 333 loci from the SBLGNT apparatus (paired McNemar $p < 10^{-3}$). Two further components we tested alongside the weighting are inert or actively harmful, so the winning configuration is the simple one — weighted Levenshtein on, everything else off. We close by deploying the system on out-of-corpus text (Photius's *Bibliotheca*) and discussing the limits of the evaluation.

1 Introduction

Ancient Greek literature survives almost entirely through chains of hand-copied manuscripts. Across centuries of transmission, copyists introduced errors: letters smudged or were misread, words were dropped or duplicated, and — because much copying was done from dictation — sounds were misheard and respelled. Recovering the text an author wrote, from witnesses that disagree with one another, is the business of textual criticism.

Logion [1] brings a language model to bear on this task. A BERT model fine-tuned on premodern Greek, it reads a passage, flags positions where the transmitted word looks improbable in context, and proposes the word it believes belongs there. Editors treat it as a tireless second reader: a way to surface suspicious readings at scale and generate candidate emendations for expert judgement.

This paper concerns one component of Logion's pipeline — how it decides which alternative spellings are "close enough" to a transmitted word to be worth scoring — and shows that a small, linguistically motivated change there produces a measurable, statistically significant improvement on the errors that matter most: the ones real scribes actually made.

2 Background and context

2.1 How Logion surfaces a suspect reading

For each word in a passage, Logion computes what its authors call a *chance-correction ratio*: how probable is the word actually written, compared with the most probable near-spelling the model would substitute in the same context? When some neighbouring spelling is far more probable than the transmitted form, that position is flagged for review and the high-probability neighbour is offered as the leading conjecture. Performance is reported as *top-k* recovery: the fraction of test loci for which the correct reading appears among the model's k highest-ranked candidates.

Computing that ratio requires a set of candidate spellings to score, which in turn requires a notion of which words count as close to the transmitted form. For this Logion uses **edit distance** (Levenshtein): the minimum number of single-character insertions, deletions, or substitutions that turn one string into another. In the public configuration the neighbourhood is every vocabulary word within edit distance 1.

2.2 Why textual criticism cares about the metric

Centuries of editorial practice have catalogued *how* Greek scribes err, and the errors are emphatically not uniform across the alphabet. A short list of confusions dominates the record:

- **Itacism.** Over the post-classical period the vowels and digraphs η, ι, υ, ει, οι all converged on the single sound /i/. A scribe taking dictation — or even one reading silently in a phonology where these were homophones — swapped them constantly. Itacism is the single largest source of orthographic variation in the manuscript tradition.
- **Visual confusions in minuscule script.** Once Greek was written in the cursive minuscule hand, certain letter shapes became easy to mistake for one another: γ/τ, λ/μ, β/ν, ρ/π. These errors are graphical, not phonetic, and follow a different distribution from itacism.
- **Post-vocalic** β/υ. A sound shift in the realisation of β after vowels bled into spelling, producing a further recurrent interchange.

A metric that assigns the same cost to η→ι (a textbook itacism) as to η→ξ (an interchange essentially never seen) discards exactly the prior knowledge editors rely on. The weighting scheme below puts that prior back into the model's candidate generation.

3 Method

3.1 From a binary gate to a weighted neighbourhood

Classical edit distance is blunt: changing one letter to *any* other letter costs exactly one. Logion's filter is blunter still — a yes/no gate that admits every vocabulary word within distance 1 and treats them all as equally plausible neighbours.

We replace the gate with a *weighted* Levenshtein neighbourhood. Each single-character substitution is assigned a cost: pairs drawn from a documented confusion class cost little, while unrelated pairs cost the full unit. The costs come from the standard reference literature on Greek pronunciation and scribal transmission — Allen's *Vox Graeca*, West's *Textual Criticism and Editorial Technique*, and Reynolds & Wilson's *Scribes & Scholars*. The filter thus becomes a continuous weight rather than a hard cutoff: candidates that differ from the transmitted word by a likely scribal slip are promoted over those that differ by an equally short but implausible edit.

3.2 A worked example

Consider the dative articles ταις and τοις, which differ only in the itacistic interchange αι→οι (both realised /i/-coloured in the relevant period). Under our scheme that substitution costs 0.35. An unrelated single-character edit of the same raw Levenshtein distance — for instance τις versus οις, where τ→ο is not a recognised confusion — costs the full 1.00. Because candidate weight falls off with cost, the itacistic neighbour receives roughly 3.7× the weight of the implausible one, even though plain Levenshtein would rank them identically. Across a large vocabulary, this re-weighting reshapes the candidate ranking toward the kinds of errors scribes actually commit.

4 Results

4.1 Real attested variants (SBLGNT)

The fairest test of a scribal-error model uses errors scribes actually made. We therefore evaluate on **real attested variants**: 333 loci in the SBL Greek New Testament apparatus where the surviving manuscripts genuinely disagree. Each locus is a place where the tradition records a competing reading; the question is whether Logion ranks the attested alternative highly when shown the passage.

On this benchmark the weighted neighbourhood recovers more variants than the baseline at every cutoff (Figure 1). The headline result is at top-5:

> **Top-5 recovery:** 42.3% **vs.** 35.7% **baseline**, a paired comparison on the same 333 loci. By McNemar's test the difference is significant at $p < 10^{-3}$: the weighted filter recovered 28 loci the baseline missed while losing only 6. Top-1 trends the same way (16.5% vs. 14.4%) but is not significant ($p = 0.30$); top-10 is 53.5% vs. 50.2%.

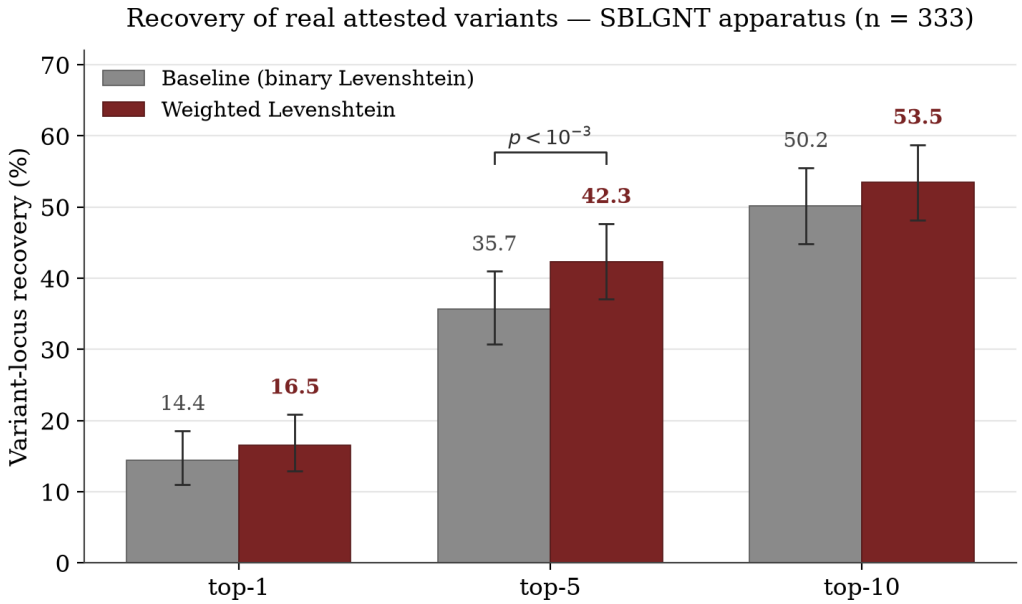

Figure 1: Recovery of real attested variants on the SBLGNT apparatus ($n = 333$), baseline (binary Levenshtein) versus the recommended weighted Levenshtein, at top-1, top-5, and top-10. The weighted filter is higher at every cutoff; the top-5 gap is significant (paired McNemar $p < 10^{-3}$). Error bars are Wilson 95% confidence intervals.

The same directional effect appears under two further, more controlled protocols. On the synthetic single-character protocol used by the original paper ($n = 200$), removing the weighted filter costs −4.0 percentage points at top-1 (one-tailed McNemar $p = 0.029$). On a corruption protocol built to mimic scribal-*class* errors rather than uniform noise, removing it costs −4.0 points at top-10 ($p = 0.039$). Three independent protocols thus agree on the sign of the effect, and the one built from genuine manuscript disagreement gives the strongest, cleanest signal.

4.2 What each component contributes

The weighted distance was developed alongside two other candidate improvements, and a single-ablation study isolates each one's contribution. Both extra components turn out to be dead weight, and — importantly — the *all-improvements-on* configuration does *worse* than baseline (−1.5pp top-1, −3.5pp top-5). Figure 2 and Table 1 summarise the per-component deltas.

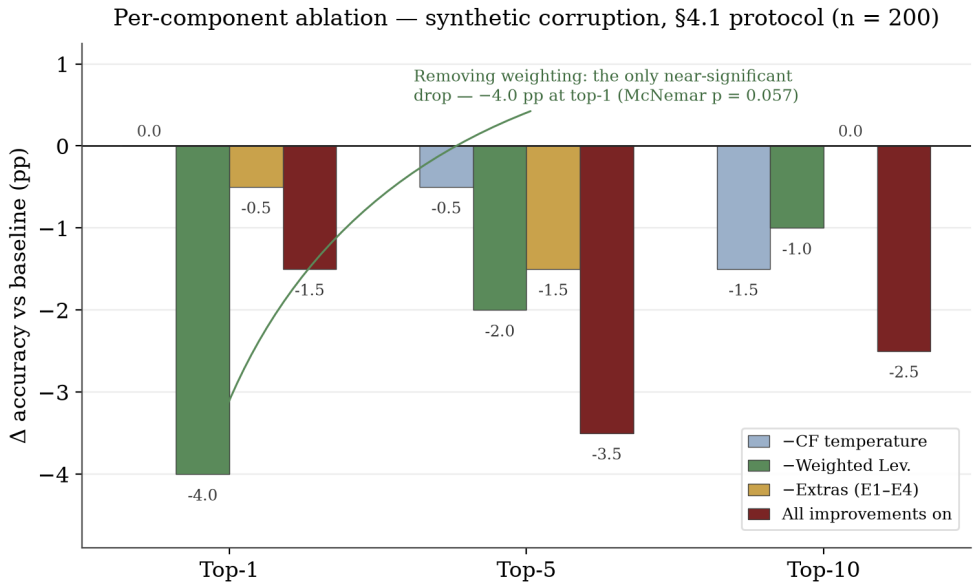

Figure 2: Per-component ablation on the synthetic single-character corruption protocol ($n = 200$, §4.1 methodology). Bars show the change in recovery accuracy relative to baseline when each component is removed (or all are turned on). Only the removal of weighted Levenshtein produces a near-significant drop.

Table 1: Per-component ablation on artificial single-character corruption ($n = 200$, §4.1 protocol). Each row removes a single component from the baseline (except the last, which enables all). p-values are paired McNemar against the baseline at top-1.

Condition	Top-1	Top-5	Top-10	McNemar p
Baseline (paper)	37.5%	69.5%	81.5%	—
− CF temperature	37.5%	69.0%	80.0%	1.00
− **Weighted Lev.**	**33.5**%	67.5%	80.5%	**0.057**
− Extras (E1–E4)	37.0%	68.0%	81.5%	1.00
All improvements on	36.0%	66.0%	79.0%	0.63

Two findings drive the recommendation:

- A **chance-fit adaptive temperature** — which flattens the model's output distribution when the written word is implausible — is inert. It ties baseline exactly at top-1 (McNemar $p = 1.00$) and moves nothing meaningfully at higher cutoffs. A separate sweep confirmed this: 14 of 20 temperature settings were worse than baseline and none reliably better.
- Four additional heuristics, the **"extras" E1–E4**, are actively harmful on a more realistic corruption protocol: removing them *lifts* top-1 from 26.5% to 30.5% ($p = 0.021$). The likely culprit is the *lectio difficilior* penalty (E1), which discounts a suggested reading more common than the transmitted word. That is exactly backwards for itacism, the most frequent error class, in which a slip routinely turns a rarer spelling into a commoner one — so the penalty suppresses exactly the corrections the model should be making.

The winning configuration is therefore the simple one: **weighted Levenshtein on, everything else off.** All headline results above use this configuration.

5 Out-of-corpus deployment: Photius

To probe behaviour beyond a held-out test set, we ran the recommended configuration over 26 paragraphs (approximately 2,100 words) of Photius's *Bibliotheca* — codices 70, 92, 161, 165, and 213 — a ninth-century text that lies outside Logion's training corpus and survives in two manuscript families (Marcianus gr. 450 and 451). The model produces tiered emendation shortlists, each candidate annotated with the orthographic operation it represents.

The shortlist behaves sensibly in aggregate: its picks cluster in the itacism and visual-confusion regions, exactly where Byzantine scribal error concentrates. None of the individual picks, however, align with documented loci in the critical apparatus for these codices. The honest reading is that the shortlist marks where the model and the transmitted text most diverge — a useful triage signal — rather than a curated set of validated emendations. This is a meaningful negative for the deployment as an *editorial* tool, discussed further below.

6 Discussion

The central finding is narrow but holds up: encoding the empirical, non-uniform structure of Greek scribal error into the candidate-generation metric improves recovery of real variants, and does so with the simplest possible intervention — a fixed cost matrix at inference time, no retraining. The gain ($\sim$6.6pp at top-5) is modest in absolute terms but consistent in direction across three protocols and significant on the only one built from genuine manuscript disagreement.

Equally informative is what *failed*. The adaptive temperature and the four extra heuristics were plausible a priori, yet the data are unambiguous: they do not help, and the *lectio difficilior* penalty actively hurts on realistic errors. The lesson is that a single, well-motivated change aligned with the domain beats a stack of clever-sounding ones, and that the all-on configuration — the tempting default — is the wrong choice here.

7 Limitations

- **Domain shift.** The public Logion checkpoint is fine-tuned on Psellus (eleventh-century Byzantine). Our variant test uses Koine of the first century and the Photius run is ninth-century, so absolute accuracy sits well below the paper's in-domain figure of 90.5%. We therefore lean on the *relative* deltas — baseline versus weighted on the same loci — which survive this shift, rather than on absolute recovery rates.
- **The synthetic protocol is biased against weighting.** The §4.1 protocol draws corruptions uniformly across the 24-letter alphabet, which quietly rewards an algorithm that treats every edit as equally likely. Since the whole purpose of weighting is to exploit the *non*-uniformity of real scribal error, this protocol understates the benefit; the real-variant test should carry the most weight.
- **A variant is not always an error.** Many SBLGNT apparatus entries record disagreements between defensible readings rather than slips of the pen. The benchmark measures whether Logion finds a locus *unusual* — a weaker claim than finding it *wrong*. Recovery of an attested variant is evidence of sensitivity to real textual instability, not proof of correction.
- **No apparatus-level validation on Photius.** The deployment shortlist clusters in the right error regions but matches no documented apparatus locus, so it cannot yet be presented as a set of emendations an editor would adopt.
- **The weights are hand-tuned.** The cost matrix is read off textbooks rather than learned. The natural next steps are to estimate the phonetic and graphical priors directly from Byzantine palaeographic corpora, and — more ambitiously — to retrain the underlying model

with the weighted neighbourhood as a *training* signal rather than only an inference-time filter, which could internalise scribal-error distributions and yield a larger gain. The code also supports multi-token emendation, which was not part of this evaluation and is a natural target for future benchmarking.

Acknowledgements

This work builds directly on Logion [1]. The SBLGNT apparatus used for the variant benchmark is from `jjmccollum/sblgnt-tei` (CC BY 4.0).

References

[1] C. Cowen-Breen, C. Brooks, B. Graziosi, and J. Haubold. Logion: Machine-learning-based detection and correction of textual errors in Greek philology. In *Proceedings of the Ancient Language Processing Workshop (ALP), ACL*, 2023.

[2] W. S. Allen. *Vox Graeca: The Pronunciation of Classical Greek*. Cambridge University Press, 3rd edition, 1987.

[3] M. L. West. *Textual Criticism and Editorial Technique*. Teubner, 1973.

[4] L. D. Reynolds and N. G. Wilson. *Scribes and Scholars: A Guide to the Transmission of Greek and Latin Literature*. Oxford University Press, 4th edition, 2013.

[5] M. W. Holmes. *The Greek New Testament: SBL Edition*. Society of Biblical Literature, 2010. Apparatus TEI encoding: `jjmccollum/sblgnt-tei` (CC BY 4.0).