GeoAffect: A Multi-Layer Annotation Schema and Few-Shot LLM Evaluation for Geoaffective Analysis of Literary Texts

Fotini Koidaki, Stergios Chatzikyriakidis
Department of Philology, University of Crete
coidacis@gmail.com, stergios.chatzikyriakidis@uoc.gr
{Fotini Koidaki,Stergios Chatzikyriakidis}@uoc.gr

Abstract

GeoAffect is an annotation framework that has been especially developed to capture how places are emotionally framed in literary narrative. The project focuses on nineteenth-century Greek prose fiction and brings together named entity recognition with an affect schema that distinguishes experiential, appraisal, and identity-oriented relations to place. The annotation design linked entities, emotion spans, and rhetorical devices, allowing us to model not only sentiment but also forms of belonging, alienation, and longing. To test the schema, we created a manually annotated gold dataset of approximately 360 sentences and evaluated thirteen Large Language Models in a few-shot setting for both entity recognition and affect classification. The results indicate that, with carefully designed prompts and selection strategies, LLMs can support structured geoaffective annotation even in low-resource historical language contexts.

Keywords: geoaffective annotation, semantic annotation schema, named entity recognition, emotion classification, large language models, few-shot learning, computational literary analysis

1. Introduction

Computational literary studies have shown growing interest in both emotion and place over the last decade — but the relationship between the two remains underdeveloped, particularly for non-English, low-resource historical corpora. The corpus of 19th-century Greek fiction from the Ionian Islands is a case in point. This corpus has only recently come to light since modern Greek literary studies have for a long time focused on Ionian poetry.[1] The Ionian authors came from a region shaped by centuries of Venetian rule, where Greek, Italian, and European cultural presences have long coexisted. A foundational question for this literary tradition remains open: do these texts share the regional identity of the Ionian Islands, the so-called *Heptanese*, or merely a shared provenance? Place and identity are closely bound (Tuan, 1977; Hall, 1996), which means that how place is emotionally framed in these texts is not a marginal question. GeoAffect was built to investigate exactly this — though the full corpus analysis lies ahead. This paper presents the GeoAffect framework and its evaluation against a manually annotated gold standard.

Existing computational approaches to emotion in literary texts rely on psychological taxonomies (Ekman, Plutchik) or dimensional models such as valence-arousal, developed to capture individual affective states in response to events or persons (Kim and Klinger, 2019). Even the most closely related work - studies that relate toponyms with sentiment in historical fiction (Heuser et al., 2016)- operates with binary or coarse-grained polarity. These approaches fail to capture the identity dimension of place-affect relations: the sense of belonging, absence of belonging, and longing to belong that characterizes how people relate to geographical space. Crucially, in literature the affective charge of a place reference is rarely expressed directly. It is constructed through narrative voice, rhetorical figuration, and implicit cultural associations. Existing tools — whether categorical emotion models or lexical resources — are not designed to capture this. To our knowledge, no annotation schema or evaluation framework exists for geoaffective classification in literary texts — let alone for a low-resource language such as 19th-century Greek. GeoAffect is our attempt to address this. We propose a multi-layer annotation schema grounded in a three-level affect taxonomy, which we evaluate against a manually annotated gold standard through systematic few-shot prompting of thirteen Large Language Models.

[1] On the relative neglect of Ionian prose fiction in favour of poetry, see Tziovas (2017), p. 83.

2. Related Work

Heuser, Moretti, and Steiner (Heuser et al., 2016) were the first to systematically map emotional associations with place names in literary texts at scale using 5,000 British novels (1700–1900), and combining named entity recognition with crowdsourced sentiment annotation. This work remains a landmark. Yet it operates within a binary polarity framework, not by choice, but because attempts to capture a broader emotional spectrum failed to produce sufficient annotator agreement.

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 68–76

This line of inquiry has been extended in recent work to incorporate other national literatures. Grisot and Herrmann (2023), for example, extend this line to German-Swiss fiction (1840–1940) through a dictionary-based approach, identifying rural and urban spatial terms via curated word lists and computing sentiment through lexicon-based scoring in fixed textual windows. Karlińska et al. (2022) examine Polish fiction (1864–1939) by tracking the emotional valence of broad spatial concepts such as "city" and "country." Together, these studies demonstrate how spatial references can be linked to affect at scale, they rely on lexicon-driven polarity measures. Our work builds on this direction by introducing a span-based annotation framework that captures the relational and identity-oriented dimensions of place in narrative.

Computational emotion research has followed mainly three approaches: categorical taxonomies such as Ekman's (1992) or Plutchik's (2001), which label discrete emotions; dimensional models that describe affect in terms of valence and arousal; and appraisal-based approaches, which explain emotion as a result of how –individuals evaluate events. These approaches are well suited to modeling clearly expressed emotional states and evaluations, and we draw on this insight in structuring our affect layer. However, they are less equipped to account for how emotion becomes attached to place as a marker of belonging or identity. In literary prose, such relations often emerge indirectly, through narrative voice or figurative language rather than explicit emotion terms.

Troiano et al. (2023) introduced an annotation schema that was grounded on appraisal theories for emotion analysis and proved that appraisals can be reliably inferred from text and improve emotion classification. However, the focus remains on event-based evaluation and does not address how places function as objects of attachment or identity within narrative. Bozia et al. (2024), studying ancient Greek and Latin corpora with a focus on identity and belonging, also draw attention to the role of place. Their analysis, however, works largely with polarity distinctions, as no specific schema is introduced for modelling place-related affect.

Large language models have recently been tested as annotation tools, with few-shot prompting approaching human performance on some classification tasks (Ziems et al., 2024). Performance becomes less reliable, however, when the task requires fine-grained theoretical distinctions: Imamovic et al. (2024) find that ChatGPT achieves high precision on Appraisal Theory labels but poor recall, with systematic errors at the level of fine-grained distinctions — a result that points to the sensitivity of theory-driven annotation to prompt design. Historical NER adds a further layer of difficulty: diachronic language variation, orthographic instability, and scarce annotated data compound extraction errors in ways that standard benchmarks do not capture (Ehrmann et al., 2023) — and 19th-century Greek prose sits squarely within these conditions. No existing work addresses all three challenges together.

3. The GeoAffect Schema

GeoAffect is a multi-layer annotation schema designed to capture the affective encoding of place references in 19th-century literary narrative. Rather than applying generic sentiment analysis to spatial entities, the schema operationalizes place-related affect as a structured, hierarchical phenomenon grounded in distinct theoretical traditions. It comprises four interdependent layers visible in Figure 1: (i) named entity recognition, (ii) emotion classification, (iii) rhetorical device annotation, and (iv) relational linking.

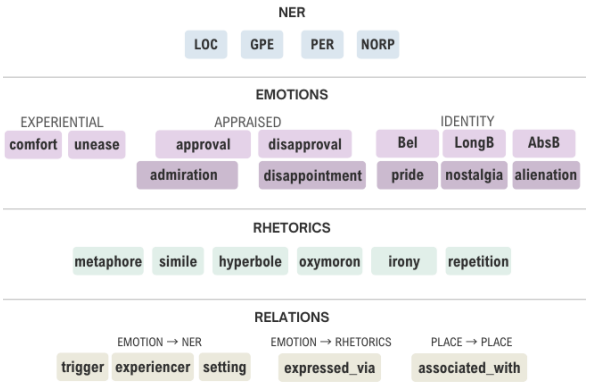

Figure 1: The four layers of the GeoAffect schema.

3.1. Named Entity Layer

The NER layer targets four entity types: locations (LOC), geopolitical entities (GPE), persons (PER), and nationalities, or religious groups (NORP). The LOC/GPE distinction ensures annotation precision for place references, separating physical or geographical spaces from entities defined by political and administrative boundaries. Entity types are annotated on the basis of their referent rather than their surface form, using adjectival or toponymic modifiers as anchors for entity identification when no proper name is present. An expression such as "the Greek state" (*elliniko kratos*, lit. "the Greek-ADJ state") is tagged GPE because it denotes a geopolitical entity, while "the lake of Kastoria" (*limni Kastorias*) is tagged LOC regardless of whether a proper name is present. The same principle applies to NORP: a collective reference such as "Italian nation" (*italikos laos*, lit. "Italian-ADJ nation") is tagged

based on what it denotes, not on whether a proper ethnonym is used.

NORP entities are treated as full-fledged targets of emotion annotation alongside place references: ethnic, national, and religious collectivities are not merely the backdrop of identity expression but one of its primary sites - the groups through which subjects define who they are, who they are not, and where they belong. This is particularly salient in the case of the Ionian (Heptanese) corpus, where collective identity is articulated at the intersection of Greek, Venetian, and Italian presences, and where belonging to a people and belonging to a place are frequently co-constructed in the same utterance.[2]

3.2. Emotion Layer

The emotion layer organizes place-related affect into three hierarchical levels. The guiding principle is that the relationship between a subject and a place is constituted experientially, evaluated culturally, and transformed into an identity narrative - each level presupposing the one below it.

The first level, ExperientialAffect, captures the immediate, pre-reflective affective tone of a place encounter: comfort and unease. Grounded in Tuan's phenomenology of place (Tuan, 1977), this level registers the bodily and sensory attunement through which a narrative subject inhabits a space before evaluation or reflection intervenes.

The second level, AppraisedAffect, captures the evaluative stance toward place as a sociopolitical and cultural space: approval and disapproval, with more intense children admiration and disappointment. Grounded in appraisal theory (Lazarus, 1991), this level treats emotion as cognitive evaluation - the judgment of a place as worthy, degraded, or disappointing within a social and ideological field. The parent/child structure reflects the intensity gradient central to appraisal models: approval intensifies into admiration when the stance takes on an aesthetic or moral dimension; disapproval deepens into disappointment when a gap between expectation and reality is affectively marked.

The third level, IdentityAffect, captures the relational orientation of a narrative subject toward place as a constitutive dimension of selfhood: belonging (Bel) with child pride, absence of belonging (AbsB) with child alienation, and longing for belonging (LongB) with child nostalgia. This level draws on Hall's account of identity as constructed through difference and exclusion (Hall, 1996) - belonging is only legible against the possibility of its absence - and on Ricoeur's theory of narrative identity, in which the self is constituted through the act of narrating experience in space and time (Ricoeur, 1992). A subject does not merely feel at ease in or approve of a place; they may recognise themselves in it, find themselves excluded from it, or orient themselves toward it as a lost or longed-for anchor of selfhood. In this sense, the three levels form a progression: a place is first experienced, then evaluated, and eventually becomes part of how the self is narrated and understood.

- **Experiential:**
  comfort (+)
  unease (–)
- **Appraised:**
  approval → admiration (+)
  disapproval →disappointment (–)
- **Identity:**
  Bel → pride (+)
  AbsB → alienation (–)
  LongB → nostalgia (+/–)

3.3. Rhetorical Layer

The rhetorical layer is a dependent sublayer: it is activated only when an emotion span has already been identified, and annotates the expressive mechanism through which affect is conveyed. The six devices recognised - irony, simile, metaphor, hyperbole, repetition, and oxymoron - were selected on the basis of their attested frequency in the corpus and their capacity to mediate place-related affect in nineteenth-century literary prose, where emotional meaning is frequently constructed through figuration rather than direct lexical expression. The layer is designed to support analysis of the relationship between rhetorical choice and affective register, without presupposing a fixed inventory of devices.

3.4. Relations Layer

The relations layer maps the structural dependencies between annotated spans. Three emotion-to-entity relations are defined: experiencer identifies the entity that undergoes the emotion; trigger identifies the entity that causally provokes it; and setting marks entities that function as background context - they spatially or socially locate the emotional expression without participating in it as agent or recipient. Any entity type (LOC, GPE, PER, NORP) may occupy any of these three roles. An expressed_via relation links emotion spans to rhetorical device spans. An associated_with

[2] Pronominal and anaphoric references to place are not currently tracked. Coreference resolution tools for 19th-century Greek are not available, which makes systematic annotation of such references impractical at this stage.

relation links entity spatial spans to each other, capturing co-reference and geographical association between named entities.

3.5. Annotation example

The following sentence illustrates how the schema captures identity-oriented place affect in practice. In this example, drawn from a narrative about Venetian nobles who had settled in Crete, two place references appear in the same sentence carrying distinct identity labels: BEL marks an active, present-tense relation to place, while NOSTALGIA marks one that has been displaced into memory. The relations layer, in this case, captures the entities that act as triggers for each emotion.

> *"Crete having become the common homeland for most of the nobles, Italy was consigned to the land of genealogical memories."*[3]

NER:
`Crete`: GPE | `Italy`: GPE

EMOTIONS:
`common homeland`→BEL | `land of genealogical memories`→NOSTALGIA

RELATIONS:
BEL $\overset{\text{trigger}}{\rightarrow}$ `Crete` | NOSTALGIA $\overset{\text{trigger}}{\rightarrow}$ `Italy`

To assign relations, we asked three questions for each annotated emotion span: what triggers it, who experiences it, and where it is located. When the answer was a named entity already marked in the NER layer, the corresponding relation was encoded — which is why in our representation relations are directed from the emotion span outward. The `associated_with` relation was reserved for cases where two spans refer to different names for the same place.

4. Pipeline & Evaluation Methodology

4.1. Gold Standard Dataset

The gold standard dataset comprises 363 manually annotated sentences drawn from a corpus of 19th-century Ionian prose fiction spanning approximately 80 texts and 11,695 pages. Very few of these were available in digital or scanned form. Source texts were digitised using a Transkribus model trained specifically on the corpus (19th-century Greek 8.0) — rather than an off-the-shelf OCR solution — in order to preserve the linguistic particularity and typographic conventions of each text, which will be relevant for future analyses.

The digitized texts were subsequently normalised from polytonic to monotonic Greek to reduce OCR error rates. After sentence segmentation, sentences were filtered for the presence of spatial entities either `LOC` or `GPE` references - using a combination of a domain-adapted spaCy model (`el_core_news_lg`) fine-tuned for this purpose,[4] and manual inspection across randomly selected texts from the Ionian corpus. The resulting sentences were stored in JSON format without any annotation and subsequently annotated in Label Studio using the full GeoAffect schema.

Annotation was carried out in two sequential phases. [5] In the first phase, all four named entity types were annotated. In the second phase, emotion spans, rhetorical devices, and relational links were annotated. The six rhetorical devices included in the schema were not selected a priori from a theoretical inventory but emerged inductively from the corpus during this phase, reflecting the devices actually attested in the material. Across the 363 sentences, annotation yielded 1,410 NER spans (GPE: 779, NORP: 277, PER: 257, LOC: 97), 836 emotion spans, and 728 relational links. Emotion labels are unevenly distributed: disapproval and unease account for over 40% of all emotion annotations, while identity-tier labels such as alienation (7) and nostalgia (11) are rare — a distribution that reflects the corpus rather than annotation bias, and that directly shaped our few-shot pool design.

Regarding span overlap, a non-overlap policy was applied within each layer: NER spans do not overlap with other NER spans, and emotion spans do not overlap with other emotion spans. NER and emotion spans likewise do not overlap with each other, as their function is relational rather than co-referential: the two layers are designed to enter into structured relations (`experiencer`, `trigger`, `setting`) rather than to describe the same textual unit simultaneously. Rare exceptions occur in cases of literary creative language where the narrative mode renders a place reference itself an affective expression. In the sentence "*It was finally printed in the Bulletin of the Medical Academy of Paris… of Paris!*", the repeated toponym is anno-

[3] Translated from a gold standard sentence. The original is written in formal 19th-century Greek prose and is considerably denser than the translation suggests.

[4] The domain-adapted `el_core_news_lg` model was fine-tuned on 150 annotated samples. It performed adequately on simple toponyms but did not generalise to compound spans, shared-head constructions, or the `GPE`/`LOC` distinction. For this reason it was used only for sentence filtering, not for entity annotation. `NORP` entities — which in this corpus include adjectival ethnic references — were excluded from its scope entirely.

[5] Annotation was carried out by a single annotator with expertise in 19th-century Greek literature and computational linguistics. Sentences where label assignment was uncertain were excluded from the test sets, which were drawn exclusively from cases of high annotation confidence.

tated as a GPE span; the repetition of this toponym functions simultaneously as a repetition span and an admiration span, since the rhetorical device is itself the vehicle of the affective expression.

4.2. Test Set Construction

Few-shot evaluation was conducted using manually curated test sets as part of a prompt optimization process aimed at producing frozen prompt configurations for subsequent large-scale corpus analysis. Random sampling from a dataset with rare labels and compound annotation structures would not provide a principled basis for evaluating the model's understanding of schema-specific phenomena.

For the NER task, test samples were selected to include the most challenging cases attested in the corpus: compound entity spans, entities referenced through adjectival or toponymic modifiers without a proper name (e.g. *Italian nation*), and elliptical coordination where a shared head noun is not repeated across conjoined entities (e.g. *the people of Greece and [of] Turkey*). This design tests whether models have internalized the referent-based annotation principle described in Section 3.1 rather than defaulting to surface-form heuristics.

For the geoaffective emotion task, priority was given to sentences with unambiguous emotional triggers and high annotation confidence - that is, sentences in which the annotator had no doubt about the label assignment. Since geoaffective analysis of literary texts is an understudied task with no prior benchmark, this selection criterion was adopted deliberately: establishing model performance under favourable conditions provides a meaningful baseline before evaluating harder cases.

Two test configurations were constructed. T1 comprises sentences with multiple co-occurring emotion annotations (two to six per sentence), selected to cover the full range of emotion labels in the schema, with preference given to sentences in which emotions are associated with place references rather than persons. T2 comprises sentences with a single emotion annotation, designed to isolate model behaviour in the absence of competing labels and contextual noise. In both configurations, all selected sentences contain at least one LOC, GPE, or even NORP entity to which the emotional expression is grounded.

4.3. Few-Shot Evaluation Pipeline

Models were evaluated via the OpenRouter API across thirteen[6] large language models spanning different model families and parameter scales, from lightweight open-weight models to frontier systems (see Table 1).[7]

Model	Provider	Scale
Claude 3.5 Sonnet	Anthropic	frontier
Claude Sonnet 4	Anthropic	frontier
Claude 3.5 Haiku	Anthropic	lightweight
GPT-4o	OpenAI	frontier
GPT-4o-mini	OpenAI	lightweight
Gemini 2.0 Flash	Google	lightweight
Gemma-2/3 27B[8]	Google	mid-scale
DeepSeek-V3	DeepSeek	frontier
DeepSeek-R1	DeepSeek	frontier
Llama 3.1 8B	Meta	lightweight
Llama 3.3 70B	Meta	mid-scale
Mixtral 8x7B	Mistral	mid-scale
Qwen-2.5 72B	Alibaba	mid-scale

Table 1: Models included in the evaluation.

All prompts were written in Modern Greek to avoid code-switching overhead introduced by translating 19th-century Greek text into English for processing. Temperature was set to 0.01 to minimise output variability across runs. Each model received the same system prompt loaded from an external file, followed by a fixed set of few-shot examples and the target sentence; no fine-tuning or parameter updates were performed. Few-shot examples for NER comprised 17 samples selected to cover the full typology of challenging cases described in Section 4.2. Few-shot examples for the affect task comprised 12 samples per configuration (T1 and T2), selected to provide clear geoaffective triggers across the emotion label hierarchy.[9]

4.4. Evaluation Metrics

The two annotation tasks were evaluated using different metrics, reflecting the different nature of their output spans.

For the NER task, the primary metrics are precision, recall, and F1 computed over exact-match entity spans, where both the text and the label must match. Exact match was chosen deliberately as a strict criterion: given that the NER test set was designed to stress-test boundary cases, a relaxed matching criterion would obscure precisely the er-

[6] Gemma 2 was used for NER few shot prompting, while Gemma 3 was used for SA.

[7] KriKri, a Greek-specific large language model, was initially considered for inclusion in the evaluation. However, response latency under long prompts with multiple few-shot examples rendered it impractical for systematic evaluation at this stage; it remains a candidate for future work.

[8] Gemma 2 27B was used for NER prompting and evaluation, and Gemma 3 27B was used for SA prompting

[9] Significance testing was not conducted given the small size of the test sets, which were designed for targeted evaluation rather than statistical inference.

rors that are most informative about model behavior.

For the geoaffective emotion task, three complementary metrics are reported. The primary metric is subset/superset F1 (SubF1): a predicted span is counted as a match if its text is a substring or superstring of the gold span with the same label. This relaxed criterion acknowledges that emotion span boundaries in literary prose are inherently less determinate than named entity boundaries, and that partial overlap with correct label assignment constitutes meaningful performance. Exact F1 is reported as a stricter reference point, and token-overlap F1 is reported to capture the degree of lexical overlap between predicted and gold spans regardless of boundary alignment. A hallucination rate - the proportion of predicted spans with no gold counterpart - is reported as a secondary metric across all models to track false positive generation, which is a known tendency of instruction-tuned models on open-ended span extraction tasks.

5. Results

Prompt optimisation for the NER task proceeded across eleven iterative versions, grouped here into three phases visible in Figure 2. From these versions, three major optimisation phases can be identified: schema rule elaboration (v0–v3.1), few-shot refinement (v4–v4.2), and structural reordering (v5–v5.1).

At baseline (v0), F1 scores ranged from 39.3% (Mixtral 8x7B) to 79.6% (DeepSeek-R1), with most frontier models clustering between 66% and 80%. The introduction of entity category definitions and general rules (v1–v2) produced modest but consistent gains across models, with Claude Sonnet 4 rising from 73.2% to 84.7%. The addition of span boundary rules, compound entity handling, and the shared head rule (v3) brought further improvement for most models, but the effect of the step-by-step recognition instructions added at this stage was non-monotonic: the verbose formulation (v3-lg) underperformed the minimal formulation (v3-mini) for ten out of fourteen models, with differences ranging from +1.0 to +16.5 percentage points in favour of the minimal style. DeepSeek-V3 and Qwen-2.5-72B showed the largest sensitivity to this choice (+16.5 and +15.2 respectively), while Claude 3.5 Sonnet was largely unaffected (–0.7). This finding suggests that procedural verbosity introduces instruction-following overhead that smaller or less instruction-tuned models cannot absorb. The most significant gains came from targeted few-shot replacement in v4–v4.2. Replacing four few-shot examples with samples specifically covering shared-head and compound entity cases raised Claude Sonnet 4 from 88.3% to 89.9%; adding one further compound example brought it to 94.5%; reordering the prompt components - placing span rules before general rules and the procedure last - produced the largest single-step gain, reaching 97.2%. This final configuration remained stable through v5 and v5.1.

Mixtral 8x7B showed a distinct pattern. Its score fell from 51.4% (v3-mini) to 29.6% at v4, , then to 26.7% and 19.6% in v4.1 and v4.2, before rising again to 47.1% at v5. No similar drop was observed in the other models. The change coincides with the few-shot replacement introduced at v4 suggesting that Mixtral is particularly sensitive to the specific few-shot examples used and that the replacement set was incompatible with its instruction-following behaviour.

5.1. NER Final Results

Table 2 reports precision, recall, and F1 for all models at v5.1. Claude Sonnet 4 achieves the highest F1 (97.2%, P=96.4, R=98.1), followed by DeepSeek-R1 (93.6%, P=92.7, R=94.4). Claude 3.5 Sonnet and GPT-4o are tied at 89.9%. Gemini 2.0 Flash reaches 87.0% at a cost of €0.005 per run - the most cost-efficient result in the upper tier. DeepSeek-V3 achieves 84.4% at €0.015, making it the strongest value proposition among mid-scale models. Mixtral 8x7B finishes last at 32.7%, confirming that its v4 collapse was not fully recovered.

Notably, Claude Sonnet 4 ranked fourth at v0 (73.2%), behind DeepSeek-R1 (79.6%), Claude 3.5 Haiku (77.1%), and Claude 3.5 Sonnet (75.9%). Its emergence as the top performer reflects a higher capacity to exploit structured prompt information - a property that was not predictable from baseline performance alone.

Model	F1	€	s
Claude Sonnet 4	**.972**	0.227	26.3
DeepSeek-R1	.936	0.043	270.8
Claude 3.5 Sonnet	.899	0.227	42.3
GPT-4o	.899	0.124	17.2
Gemini 2.0 Flash	.870	0.005	12.2
DeepSeek-V3	.844	0.015	53.6
Claude 3.5 Haiku	.833	0.063	41.3
Qwen-2.5-72B	.756	0.029	50.8
Llama 3.3-70B	.733	0.019	42.8
Llama 3.1-70B	.685	0.017	38.7
Llama 3.1-8B	.614	0.002	44.9
GPT-4o-mini	.600	0.007	23.5
Gemma-2-27B	.523	0.004	31.5
Mixtral 8x7B*	.327	0.022	30.5

Table 2: NER results at v5.1 (final prompt), sorted by F1. € = cost per run; s = avg. response time. *Mixtral collapsed at v4–v4.2; see Section 5.1.

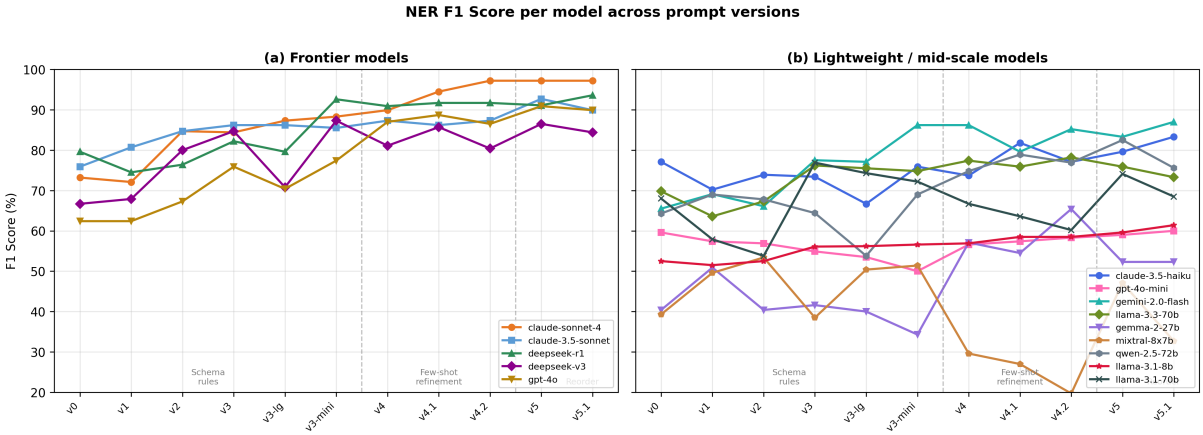

Figure 2: NER F1 score per model across prompt versions. Dashed vertical lines delimit the three optimisation phases: schema rule elaboration (v0-v3.1), few-shot refinement (v4-v4.2), and structural reordering (v5 v5.1).

5.2. Geoaffective Analysis: Test T1

Building an effective few-shot pool required several rounds of adjustment. We began with 12 examples, ensuring that each emotion label was represented at least twice, and conducted an initial evaluation to identify systematically weak categories. `LongB`, `AbsB`, and `pride` showed lower recognition rates, prompting the addition of targeted examples, first expanding the pool to 13 and then to 15 samples. All three configurations were tested across four prompt versions, ranging from a minimal schema description (v1) to a more elaborated formulation with explicit guidance for potentially confusable labels (v4).

Exploratory trials with pools exceeding 15 examples were also conducted. In these cases, performance decreased across models. Since T1 already consists of compound, multi-label sentences, increasing the number of examples appeared to add contextual load without improving discrimination. For this reason, 13 and 15 examples were treated as the upper bounds in this setting.

What we found was that adding more examples did not reliably help. For Claude Sonnet 4 and Claude 3.5 Sonnet, pool_13 gave the optimal (91.4% and 89.2% SubF1), while pool_15 resulted in lower performance for both. DeepSeek-V3 peaked at pool_12. Beyond a certain threshold, adding further examples appear to introduce noise rather than useful signal — and where that threshold falls differs by model.

What proved more important, however, was how prompt version and pool size interacted with each other. For Claude Sonnet 4, GPT-4o, and Gemini 2.0 Flash, v4 improved consistently over v1 regardless of pool size. For Claude 3.5 Sonnet, DeepSeek-V3, and DeepSeek-R1 the outcome depended heavily on which pool was used: at pool_13, v4 caused notable drops — 9.8 points for Claude 3.5 Sonnet, 13.9 for DeepSeek-R1, 18.4 for DeepSeek-V3 — while at pool_15 the same prompt helped or had negligible effect. A richer prompt does not simply add guidance; it interacts with the few-shot composition in ways that are not easy to predict.

Model	$v1_{12}$	$v1_{13}$	$v1_{15}$	$v4_{12}$	$v4_{13}$	$v4_{15}$
Consistent improvement with v4						
Sonnet 4	84.1	91.4	90.1	91.4	**93.2**	**93.2**
GPT-4o	69.0	72.4	66.7	73.7	67.9	**74.2**
Gemini Flash	69.8	71.9	67.7	70.8	**76.9**	**74.6**
Pool-dependent response to v4						
Sonnet 3.5	83.6	**89.2**	81.8	85.3	79.4	80.0
DeepSeek-V3	**83.9**	80.6	69.8	65.5	74.2	74.6
DeepSeek-R1	70.8	**77.4**	63.3	68.8	63.5	66.7

Table 3: SubF1 (%) for prompt versions v1 and v4 across pool sizes (12, 13, 15), Test T1. Bold: best per model.

5.3. Geoaffective Analysis: Test T2

Test T2 used the same prompts as T1, but this time each sentence carried a single emotion label — though not necessarily a single span. Running this simpler setting alongside T1 let us see where the models genuinely struggle with a label, and those per-label weaknesses were precisely what guided our few-shot pool expansions in T1.The difference in overall performance was clear: most models dropped to zero hallucination across all prompt versions, and results improved steadily from v1 to v4.

We also saw a different set of models at the top. Gemini 2.0 Flash and Claude 3.5 Sonnet reached the highest scores, with Gemini being the only

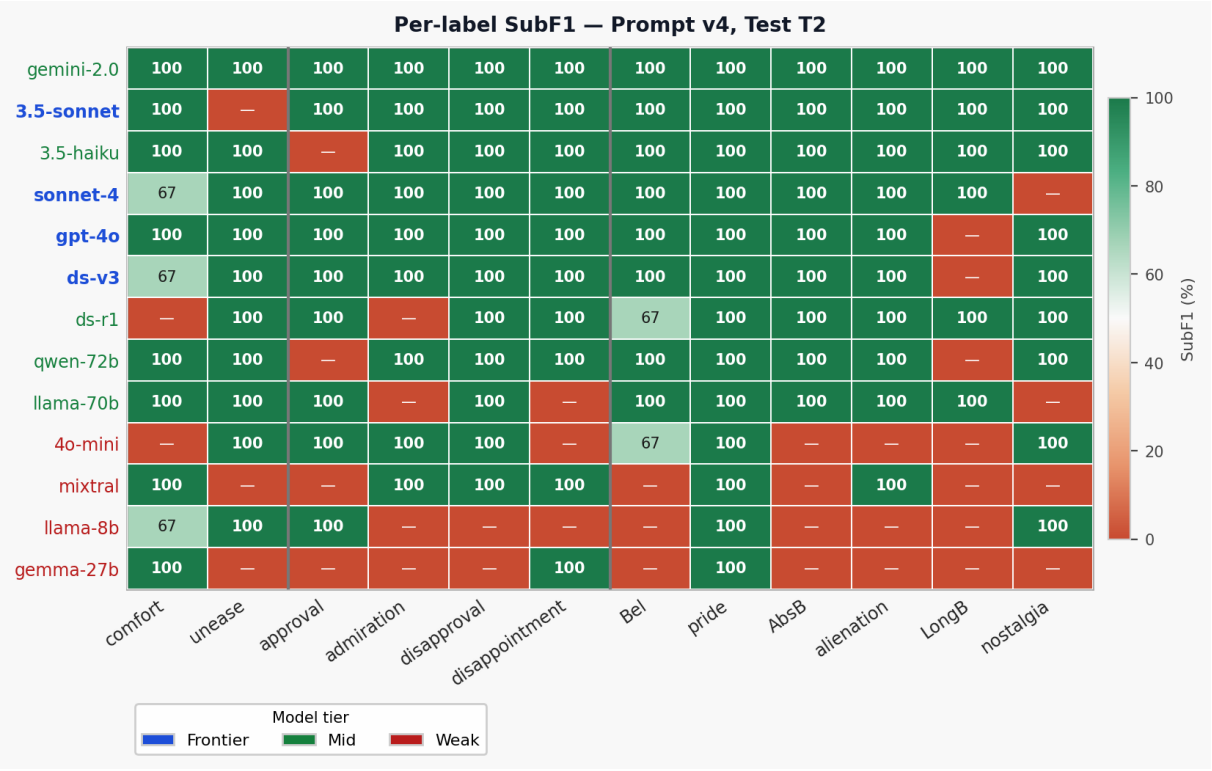

Figure 3: SubF1 per label and model — prompt v4, Test T2.

model to achieve 100% SubF1 at v4 — a result we ran multiple times to confirm. What makes this particularly interesting is the trajectory: Gemini started at 85.7% at v1 and improved consistently with each prompt version, ending at a perfect score. Its Token F1 of 84.7%, however, shows that perfect label recognition does not always mean perfect span boundaries. The model identified the right labels consistently but did not always capture the exact extent of the span. Claude 3.5 Haiku also performed well above expectations for its size. Claude Sonnet 4, our best performer in T1, came in at 92.3%. When the task is cleaner, the advantage of the largest models seems to become less pronounced.

Model	Sub F1	Tok F1	Halluc
Gemini 2.0 Flash	**1.00**	.847	.000
Claude 3.5 Sonnet	.957	.720	.000
Claude 3.5 Haiku	.957	.700	.000
Claude Sonnet 4	.917	.785	.083

Table 4: T2 results at v4, top models.

At the label level, most categories posed no particular difficulty by v4. `LongB` stood out as the hardest, given that 7 out of 13 models missed it entirely, across all versions. A second group of labels (`approval`, `admiration`, `AbsB`, `nostalgia`, and `Bel`) also showed complete misses in 3–4 models, suggesting that identity-tier and finer appraised distinctions remain genuinely hard even in a single-label setting. `comfort` was a different case: most models recognised it but some failed to capture the exact span, producing partial matches rather than complete misses.

6. Conclusions

GeoAffect was designed to study how place is affectively framed in nineteenth-century Greek prose and to test whether such distinctions can be supported through few-shot LLM annotation.

The experiments show that performance depends less on longer or detailed prompt descriptions and more on the targeted selection of few-shot examples, which appear to be the most effective learning source for LLMs. At the same time, we observe a ceiling effect, since after about 12–15 examples, improvements tended to level off, and in some cases additional examples introduced noise. This also has practical implications, given the cost associated with longer prompts.

Model behaviour also varied in consistency. Even at low temperature, some systems produced unstable outputs. At the same time, lighter models proved competitive in simpler configurations. Gemini 2.0 Flash stood out in T2, reaching the highest SubF1, while remaining faster and cheaper than frontier models. With further prompt and example refinement, it appears suitable for scaling annota-

tion to the full Ionian corpus.

Overall, the experiments indicate that LLMs can function as annotation and information extraction tools for digital humanities tasks beyond their primary training domains, provided that the schema is clearly specified and evaluation distinguishes label recognition from span accuracy.

The GeoAffect framework is still under active development. Having a clearer picture of how models respond to different prompt configurations, our next step is to select the best candidate — in terms of accuracy, speed, and cost — and apply it to the full Ionian corpus.

7. Acknowledgements

Funded by the European Union under Horizon Europe (project TALOS-AI4SSH, G.A. 101087269)

8. Bibliographical References

Eleni Bozia, Austin Stein, Wavid Bowman, Annie Gjineci, Gillian Vilela, Zachary Hracho, Rohan Prasad, Neema Owji, Niloufar Saririan, Aidan Burrowes, Aarushi Jain, and Nitaicandra Stevens. 2024. Performing sentiment analysis to trace the history of identity and belonging in ancient greek literature. *Digital Scholarship in the Humanities*, 39(4):1019–1025.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Giulia Grisot and Berenike Herrmann. 2023. Examining the representation of landscape and its emotional value in german-swiss fiction between 1840 and 1940. *Journal of Cultural Analytics*, 8.

Stuart Hall. 1996. *Who Needs Indentity*, 3rd edition, pages 19–33. SAGE Publications LTD, London, Thaousand Oaks, New Delhi.

Ryan Heuser, Franco Moretti, and Erik Steiner. 2016. The emotions of london. *Stanford Literary Lab Pamphlets*, 13(101).

Mirela Imamovic, Silvana Deilen, Dylan Glynn, and Ekaterina Lapshinova-Koltunski. 2024. Using ChatGPT for annotation of attitude within the appraisal theory: Lessons learned. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 112–123, St. Julians, Malta. Association for Computational Linguistics.

Matthew Jockers. 2015. The ancient world in nineteenth-century fiction; or, correlating theme, geography, and sentiment in the nineteenth century literary imagination. *Digital Humanities Quarterly*, 10(2).

Agnieszka Karlińska, Cezary Rosiński, Jan Wieczorek, Patryk Hubar, Jan Kocoń, Marek Kubis, Stanisław Woźniak, Arkadiusz Margraf, and Wiktor Walentynowicz. 2022. Towards a contextualised spatial-diachronic history of literature: mapping emotional representations of the city and the country in Polish fiction from 1864 to 1939. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 115–125, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Evgeny Kim and Roman Klinger. 2019. An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

Richard S Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press, New York.

Niu Minxue, Jaiswal Mimansa, and Emily Mower Provost. 2024. From text to emotion: Unveiling the emotion annotation capabilities of llms. *ArXiv*, abs/2408.17026.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Paul Ricoeur. 1992. *Oneself as Another*. University Of Chicago Press, Chicago.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Yi-Fu Tuan. 1977. *Space and Place: The Perspective of Experience*. University of Minnesota Press, Minneapolis, MN.

Dimitris Tziovas. 2017. *I Politismiki Poiitiki tis Ellinikis Pezografias: Apo tin Ermineia stin Ithiki*. Crete University Press, Heraklio.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.