Retrieval-Augmented vs Prompt-Only Governance Architectures: Selective Effects on Register Drift in LLM Book Companions

Technical Note

Rayan B. Vasse Independent Researcher

Abstract

We compare two governance architectures for controlling register drift in LLM book companions: retrieval-augmented generation with structured Canon Packs (CPFull) versus prompt-only governance specifications (CPLite), both running on Claude Sonnet 4.0 via direct API calls. Across 8 books and 15 questions (N = 120 paired observations), adding RAG to prompt-only governance produces *category-specific* effects rather than uniform drift reduction: a significant medium-sized reduction in productivity/application drift (D2: W = 201, p = .014, r = .40), a non-significant reversal in doctrinal flattening (D3: marginally higher under RAG), and no detectable effect in therapeutic (D1) or chatty-assistant (D4) categories. The overall corpus-level effect is therefore non-significant (W = 754, p = .169), but this null masks the operationally important differential pattern. A plausible methodological objection - that lexicon-based drift instruments would falsely flag faithful retrieval as drift when source-text vocabulary overlaps with detection markers - is empirically tested and not supported (Spearman ρ = .048, p = .911 between per-book source-marker density and RAG effect). The lexicon is therefore robust against this surface-marker confound for the present corpus. The D4 category was near-zero in both Sonnet conditions, contrasting with substantial D4 drift previously observed for GPT-4o-mini, indicating that D4 behaviour is model-specific rather than governance-dependent. The two architectures compared differ on both retrieval and prompt-specificity dimensions; observed effects reflect the combined intervention and should not be attributed to retrieval in isolation.

Keywords: register drift, retrieval-augmented generation, prompt-only governance, LLM book companions, lexicon-based measurement, computational hermeneutics

1. Introduction

Register drift describes the tendency of LLM assistants to converge toward a generic, therapeutic, productivity-oriented discourse style regardless of the source text they are meant to represent (Vasse, 2025). Two architectural approaches to governing this drift have been proposed: prompt-only governance, in which behavioural constraints are injected as system-prompt instructions; and retrieval-augmented generation (RAG), in which chunked source text is retrieved and injected alongside governance instructions. A common assumption is that grounding responses in retrieved passages anchors the model in the source text's register, thereby reducing drift. This note tests that assumption empirically on the same model, the same question battery, and the same set of books.

The contribution of the note is threefold. First, it provides a matched-model comparison of two governance architectures under controlled conditions. Second, it characterises the *direction-by-category* of the RAG effect rather than reporting only a corpus-level summary statistic, which we find misleading. Third, it empirically tests and rejects a plausible methodological objection to lexicon-based drift measurement: that retrieval should inflate measured drift when source-text vocabulary overlaps with detection markers. The rejection is reported as a positive finding about instrument robustness.

A caveat that frames everything below: the two architectures we compare differ along *more than one dimension*. CPFull combines retrieval with a substantially more detailed system prompt; CPLite uses a leaner prompt and no retrieval. The observed effects therefore reflect this combined difference and should be read as a comparison of deployed architectures as-built, not as an isolated test of retrieval. We address this explicitly in §2 and §4.4.

2. Method

Conditions. Two governance conditions were compared, both using Claude Sonnet 4.0 (`claude-sonnet-4-20250514`) at temperature 0.4:

- **CPFull/Sonnet (RAG-grounded).** User queries are embedded (OpenAI `text-embedding-3-small`), the top-5 most similar chunks are retrieved from a per-book

vector store (512-token chunks, 64-token overlap), and injected as retrieval context alongside a structured Canon Pack system prompt (~4KB) containing interpretive framework, voice configuration, chapter reference, and boundary rules.

- **CPLite/Sonnet (prompt-only).** A condensed governance specification (~500 words; fidelity stance, tone, grounding instruction, may-do/avoid/red-line boundaries) is provided as the system prompt. No source text is retrieved; the model relies on parametric knowledge and governance instructions alone.

A note on what the comparison isolates. The two conditions differ on two dimensions: presence vs absence of retrieval, *and* system-prompt richness (~4KB vs ~500 words). The comparison therefore tests deployed architectures as-built; it does not isolate the retrieval contribution. A cleaner test would hold the system prompt constant and vary only retrieval. We treat the present design as an ecologically meaningful comparison of two deployed configurations rather than a controlled retrieval manipulation, and we return to this in §4.4.

Both conditions are direct Anthropic API calls; neither uses the Living Literature platform's user-facing routing layer. Consequently, the platform interface artefact described in Vasse (2025, Assistant Register submission) - where suffix-appended routing instructions generated the majority of apparent D4 markers in CPLite-platform deployments - does not affect either condition here.

Corpus. Eight books spanning philosophy, autobiography, drama, social theory, and military strategy: *The Art of War* (Sun Tzu), *Confessions* (Augustine), *Fourth Culture* (Vasse), *Autobiography* (J. S. Mill), *Narrative of the Life of Frederick Douglass*, *Meditations* (Marcus Aurelius), *Nathan the Wise* (Lessing), and *The Souls of Black Folk* (Du Bois). These 8 of a 10-book registry were selected on availability grounds: all had both a CPFull vector store (processed through the GenAI Hermeneutics pipeline) and a CPLite governance specification (produced during onboarding on the Living Literature platform). The remaining two registry titles (*Content*, Doctorow; *Being No One*, Metzinger) lacked completed governance specifications at the time of analysis. *Fourth Culture* is included here because the source text is available to the author for full-text indexing; it is excluded from the companion Assistant Register study to preserve that paper's validation independence from the H2H proof-of-concept text.

Protocol. Each condition responded to 15 questions from the H2H protocol (Vasse, 2025, Appendix B), organised in five blocks targeting different drift vulnerabilities: core fidelity (B1), reader-helpfulness boundary (B2), modern application stress (B3), belief/framework specificity (B4), and high-drift boundary stress (B5).

Measurement. Register drift was quantified using the 43-marker Tier 1 lexicon from H2H Appendix A, spanning four categories: D1 (therapeutic/self-help, 16 markers - corrected from earlier 18-count typo), D2 (productivity/application, 10 markers), D3 (doctrinal/intellectual flattening, 8 markers), and D4 (chatty-assistant/audience-smoothing, 10 markers). Marker counts were normalised per 1,000 tokens. Paired Wilcoxon signed-rank tests (two-sided) compared conditions across 120 book-question pairs. Effect size is reported as $r = Z/\sqrt{N}$.

3. Results

3.1 Overall drift. Mean Tier 1 drift was 1.73 markers per 1,000 tokens for CPLite and 1.44 for CPFull. The difference was not statistically significant ($W = 754$, $p = .169$, $r = .18$). CPFull produced lower drift than CPLite in 34 of 120 pairs (28%). Mean response length was greater for CPFull (M = 322 tokens) than CPLite (M = 247 tokens); all drift scores are normalised per 1,000 tokens.

3.2 Per-category effects (Figure 1). The aggregate null masks category-specific effects, which are the operationally important findings.

- **D2 (productivity/application):** Significantly lower in CPFull (M = 0.60) than CPLite (M = 0.97); $W = 201$, $p = .014$, $r = .40$ (medium effect). This is the only category reaching significance and is the study's primary positive finding.
- **D1 (therapeutic/self-help):** Non-significant reduction (CPLite = 0.30, CPFull = 0.26, $p = .184$).
- **D3 (doctrinal/intellectual flattening):** Trended in the opposite direction — CPFull produced marginally *more* D3 markers (CPFull = 0.55) than CPLite (0.46), though not significantly ($p = .962$). The reversal is interpreted in §4.2.
- **D4 (chatty-assistant):** Near-zero in both conditions (CPLite = 0.00, CPFull = 0.03). This floor effect is discussed in §4.3.

3.3 Per-book variation (Figure 2). The direction of the RAG effect varied substantially by book:

Book	CPLite	CPFull	Diff	Direction
Fourth Culture	1.81	0.46	+1.35	RAG reduces
Nathan the Wise	1.80	0.67	+1.13	RAG reduces
Autobiography (Mill)	3.42	2.69	+0.74	RAG reduces
Narrative (Douglass)	1.71	1.69	+0.01	Neutral
The Art of War	0.57	0.68	−0.11	RAG increases

Meditations	1.06	1.21	−0.15	RAG increases
The Souls of Black Folk	0.73	0.96	−0.23	RAG increases
Confessions	2.75	3.16	−0.40	RAG increases

Three books showed substantial RAG benefit; four showed RAG increasing measured drift; one was neutral. With only 8 book-level observations the variation is descriptive; we make no inferential claim about book-level differences.

3.4 Test of the register-overlap hypothesis (Figure 4). A plausible mechanistic interpretation of the per-book variation is that books whose source vocabulary overlaps with the drift lexicon will show inflated drift scores under RAG, because retrieval introduces lexicon-matching terms that are faithful to the source but penalised by the instrument. The hypothesis predicts a negative correlation between per-book source-text marker density and the RAG effect: more overlap → worse RAG outcome.

We computed the Tier 1 marker density in each book's source text (markers per 1,000 tokens of source) and correlated it with the per-book RAG effect (CPLite drift − CPFull drift). The hypothesis was not supported. Pearson $r = .538$ ($p = .169$); Spearman $\rho = .048$ ($p = .911$). The correlation, to the extent it exists, trends positive - the opposite direction from prediction. The book with the highest source-text marker density (*Fourth Culture*, 0.60 markers per 1,000 source tokens) showed the strongest RAG benefit (+1.35). Books with moderate density showed effects in both directions.

This finding has methodological force beyond the present study. A reasonable skeptic of lexicon-based drift measurement would predict that surface-marker overlap should inflate measured drift in retrieval-augmented conditions, producing a false-positive bias whenever the source text contains words that double as drift markers. The current data falsify this skeptical prediction for this lexicon and corpus: the surface-marker confound is not detectable. The lexicon is therefore robust against this specific class of objection, at least for the books and architecture tested here.

What the present analysis does *not* explain is the per-book variation itself, which remains real but unaccounted for by measurable text properties at the surface level. Candidate explanations include differences in Canon Pack specificity across books, differences in training-data saturation (how strongly the model has internalised each text from pre-training), or source-text argumentative structure - but none are testable with the current design.

3.5 Per-block variation (Figure 3). The RAG advantage was concentrated in blocks targeting modern application (B3: mean diff = +0.69) and belief/framework specificity (B4:

mean diff = +1.01). Core fidelity (B1: +0.02), reader-helpfulness (B2: +0.02), and high-drift stress (B5: +0.04) questions showed negligible differences. This suggests RAG grounding is most useful when questions invite the model to extrapolate beyond the text, where retrieved passages may constrain responses toward source-anchored framing rather than generic defaults.

4. Discussion

4.1 The selective D2 effect. The only significant positive finding is a medium-effect reduction in productivity/application drift under RAG ($p = .014$, $r = .40$). D2 markers - *mindset*, *empower*, *growth*, *overcome*, *inspire*, *achieve your goals* - represent generic application framing that retrieved source passages appear to displace. When the model has access to actual book content, it anchors in source-specific arguments rather than defaulting to productivity language. This is convergent with prior findings (Vasse, 2025; Assistant Register submission) that D2 is the most governable drift category across multiple governance architectures. The category that is most consistently reduced by intervention - whether the intervention is a Canon Pack governance object, retrieval-augmented context, or both - is the productivity-application register. We regard D2 as the empirically most tractable drift category in the current Tier 1 framework.

The *Mill* anomaly deserves note. Mill's *Autobiography* shows the third-largest RAG benefit (+0.74) despite a register that is not particularly distinct from drift-marker vocabulary - utilitarianism, intellectual development, and the documented mental crisis are all standard drift-adjacent themes. A speculative account: Sonnet 4.0 may have lower parametric familiarity with Mill's *Autobiography* than with Stoic or Augustinian texts that are very heavily represented in training data, so retrieval contributes more genuinely new information for Mill than for *Meditations* or *Confessions* where the model already has rich parametric knowledge. The current design cannot test this; we flag it as a candidate for future work that varies underlying-model familiarity with target texts.

4.2 The D3 reversal. CPFull produced marginally more doctrinal-flattening markers than CPLite (0.55 vs 0.46). One plausible mechanism: retrieved passages introduce bridging and framing vocabulary from the source text's own transitional structure - chapter conclusions, editorial summaries, thematic restatements - which the lexicon categorises as flattening (*timeless*, *journey*, *it's worth noting*, *meaningful journey*, *meaningful experience*). The same retrieval mechanism that displaces generic productivity framing (reducing D2) may simultaneously introduce source-derived bridging vocabulary that the lexicon counts as flattening (increasing D3). This is not a separate phenomenon but a

downstream consequence of retrieval injecting source-derived text into the response space.

The wider methodological implication is that fixed-lexicon instruments cannot always distinguish faithful source-derived register from generic drift. The *register-overlap* objection tested in §3.4 was the obvious version of this concern at the corpus level, and it was not supported. The D3 reversal suggests a subtler version of the same concern that operates at the category level: even when overall marker density is not predictive, retrieval may shift drift mass *between* categories in ways the instrument cannot disambiguate. Future revisions of the lexicon should consider distinguishing source-derived from model-default occurrences of bridging vocabulary, possibly through retrieval-aware scoring.

4.3 The D4 floor effect. Chatty-assistant markers were near zero in both conditions (CPLite = 0.00, CPFull = 0.03). Comparison with prior results on GPT-4o-mini (Vasse, 2025; Assistant Register submission) - where D4 was a substantial drift category requiring suffix-stripping intervention to interpret, with 88% of apparent D4 attributable to the platform's interface routing - suggests that the D4 category is model-specific, varying with the underlying model's post-training behaviour. Sonnet 4.0 shows a near-zero D4 baseline at the response level; GPT-4o-mini shows substantial D4 drift even in the ungoverned condition.

This contrast is compounded here by the absence of platform routing in either condition (§2), but the model-level difference is architecturally meaningful and independent of the routing question: when both models are evaluated through direct API calls without platform-side suffix appendage, Sonnet 4.0 produces near-zero D4 markers in response content while GPT-4o-mini produces substantially more. Sonnet 4.0 appears to have internalised the avoidance of overt assistant-register markers to a degree that GPT-4o-mini has not. For Sonnet, D4 is an architectural invariant rather than a governance outcome.

4.4 Limitations. The two conditions differ on more than the retrieval dimension alone. CPFull's Canon Pack system prompt (~4KB with chapter-level interpretive scaffolding) is substantially more detailed than CPLite's governance specification (~500 words). Observed differences therefore reflect the combined effect of retrieval *and* prompt specificity, and cannot be attributed to retrieval alone. A cleaner test would hold the system prompt constant and vary only the presence of retrieved context; the current design compares two deployed architectures as-built rather than isolating a single variable. We have flagged this in §2 and treat the result throughout as a comparison of two configurations rather than a controlled retrieval manipulation.

The corpus of 8 books, while spanning diverse genres, yields only 8 book-level observations — insufficient for reliable per-book inferential testing. We restrict per-book analysis to description and frame the per-book variation as observed-but-unexplained.

Sonnet 4.0 is the only model tested. Whether the same selective pattern would appear on GPT-4o-mini, Claude Opus, or open-weight models is open. The D4 floor effect is itself evidence that model identity matters substantially for drift behaviour; the corresponding D1, D2, and D3 effects may also be model-specific in ways not visible from a single-model design.

The lexicon itself is theoretically grounded but has not been benchmarked against external psycholinguistic instruments (LIWC, Empath) in the present study. Convergent-validity testing remains a separate planned analysis.

5. Conclusions

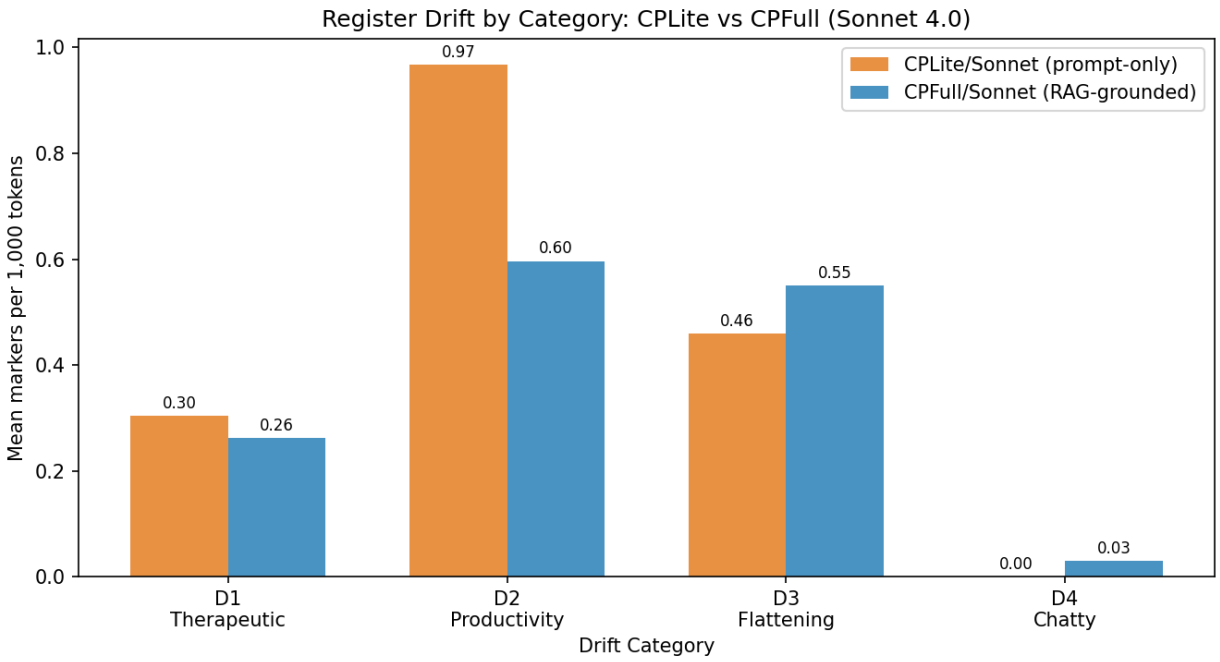
1. **Adding RAG to prompt-only governance produces category-specific effects, not uniform drift reduction.** The overall corpus-level effect is non-significant ($p = .169$), but this null masks operationally important differential patterns.
2. **RAG selectively reduces productivity drift (D2)** by a medium effect ($p = .014$, $r = .40$), replicating the cross-paper finding that D2 is the most governable drift category in the current Tier 1 framework.
3. **RAG provides no benefit for doctrinal flattening (D3),** which trends marginally higher under retrieval - plausibly due to source-derived bridging vocabulary entering the response through retrieved passages.
4. **The D4 (chatty-assistant) category is model-specific,** near-zero for Sonnet 4.0 regardless of governance architecture, in contrast with substantial D4 drift previously observed for GPT-4o-mini. This is a model-architectural finding about post-training behaviour rather than a governance outcome.
5. **The per-book variation is real but not predicted by source-text marker density.** The lexicon is therefore robust against the surface-marker confound that a methodological skeptic might most plausibly invoke against lexicon-based drift measurement on register-overlapping texts. We report this as a positive finding about instrument robustness.
6. **Prompt-only governance achieves comparable drift control for therapeutic (D1) and chatty-assistant (D4) categories;** RAG provides selective benefit for productivity drift (D2) but no benefit for doctrinal flattening (D3). The engineering overhead of chunking, embedding, and retrieval is therefore justified on register-

fidelity grounds only for productivity drift specifically, not as a general drift-reduction solution. The case for RAG in book companions must rest on other benefits - factual grounding, quotation accuracy, passage-level reader engagement - rather than register governance alone.

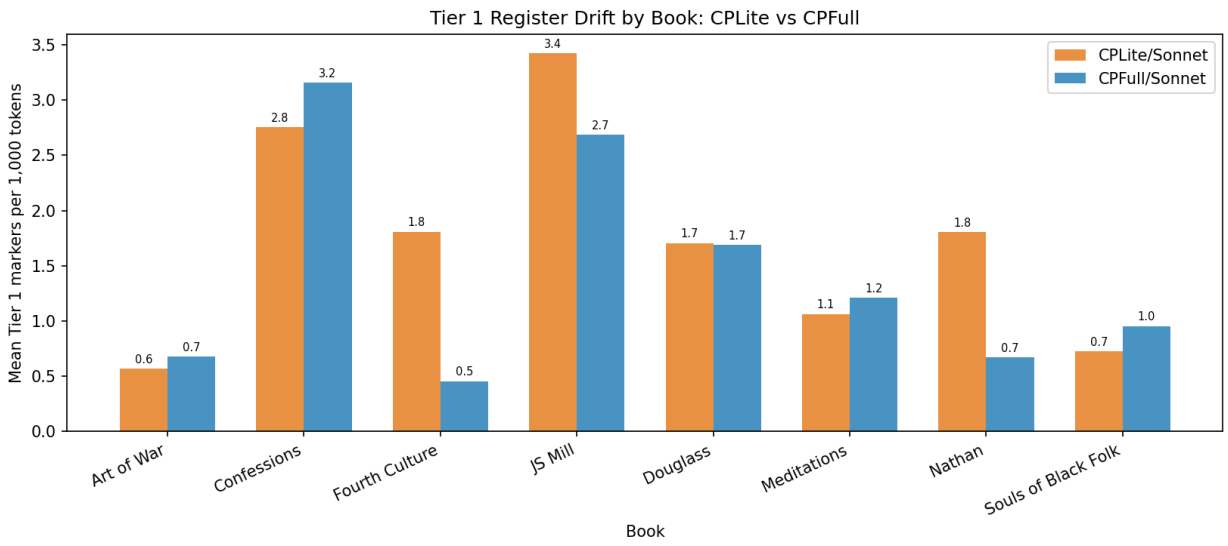
7. **The comparison reported here is between deployed architectures as-built, not a controlled isolation of retrieval.** Observed effects reflect the combined contribution of retrieval and the richer system-prompt scaffolding of the CPFull condition. Disentangling these requires a follow-up design holding prompt richness constant.

Figures

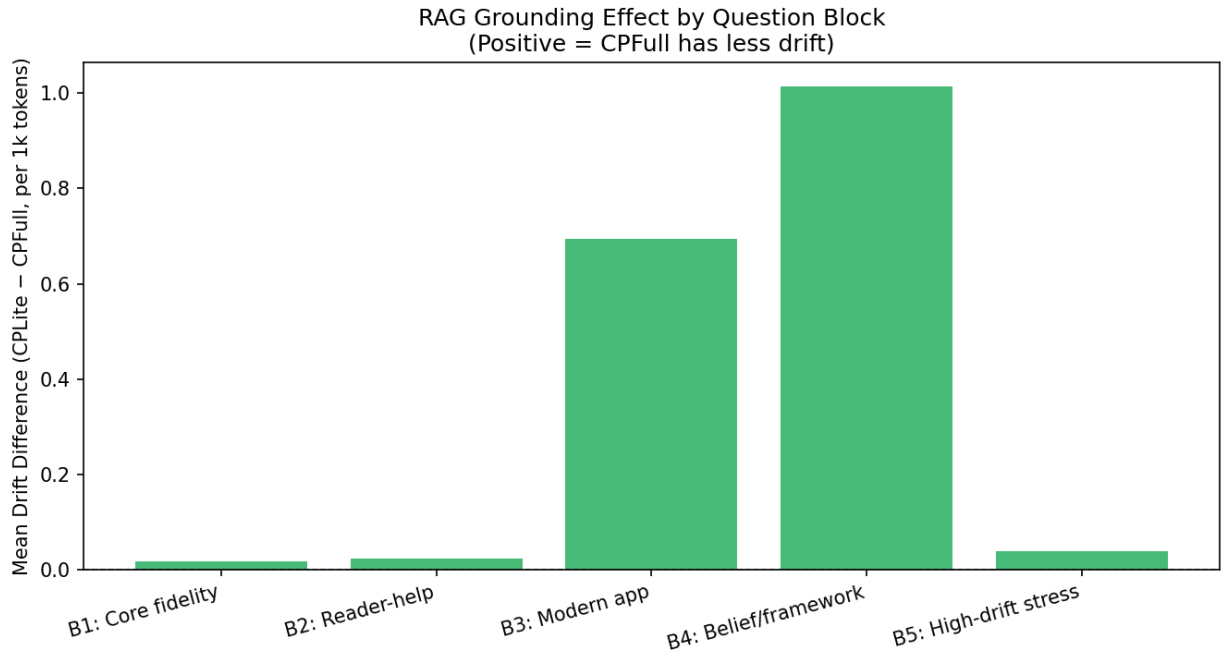
- **Figure 1.** Register drift by category (D1–D4): CPLite/Sonnet vs CPFull/Sonnet. D2 (productivity) is the only category with significant reduction under RAG (p = .014). D3 (flattening) trends in the opposite direction. D4 is near zero in both conditions. *File: chart_categories.png*



- **Figure 2.** Tier 1 register drift by book. The RAG effect varies in direction across books and is not predicted by source-text marker density (see Figure 4). *File: chart_by_book.png*

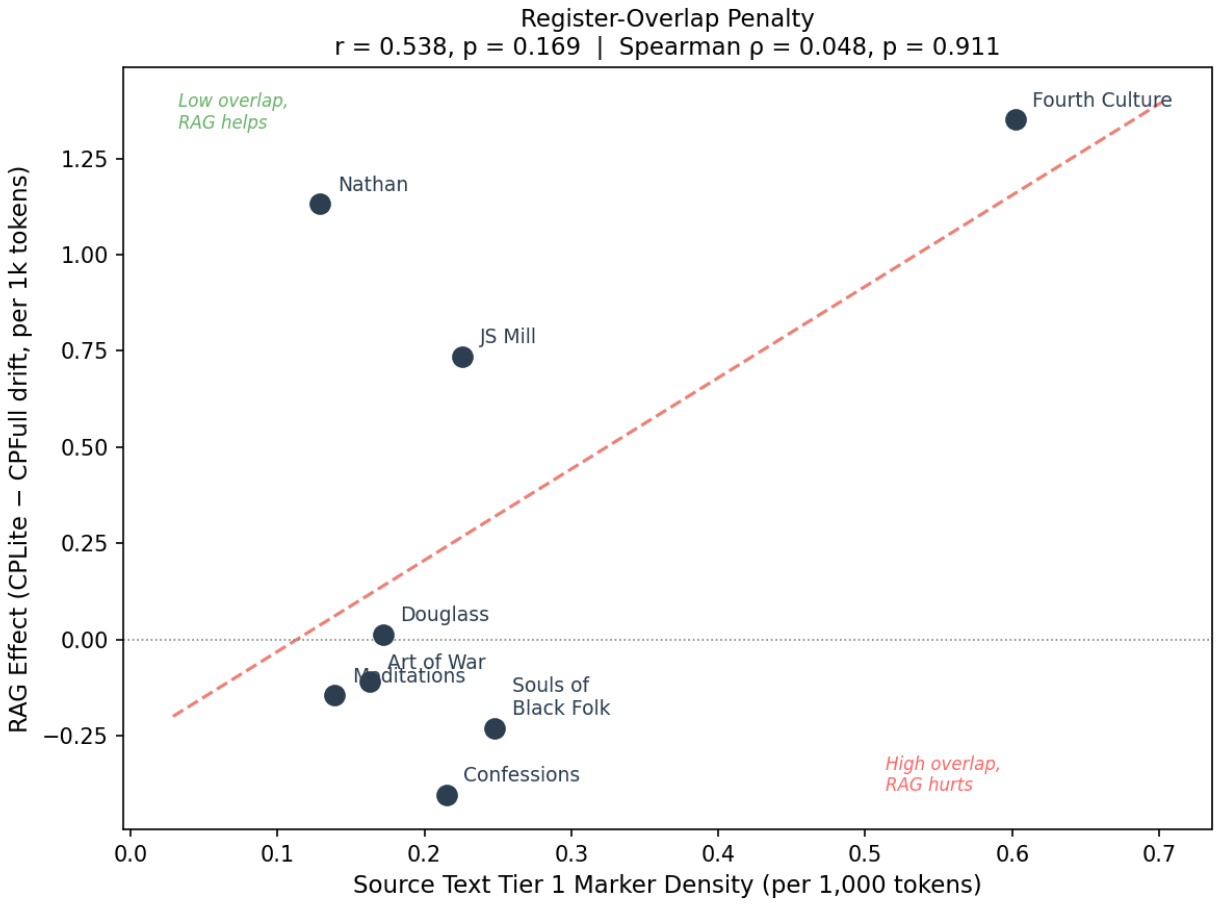


- **Figure 3.** RAG grounding effect by question block. Positive values indicate CPFull has less drift. The advantage is concentrated in Blocks 3 (modern application) and 4 (belief/framework specificity), where questions invite extrapolation beyond the source text. *File: chart_by_block.png*



- **Figure 4.** Source-text Tier 1 marker density (x-axis) vs RAG effect on measured drift (y-axis), per book. Pearson r = .538, p = .169; Spearman ρ = .048, p = .911. The register-overlap hypothesis - that higher source-text marker density should predict

worse RAG outcomes - is not supported. The lexicon shows robustness against the surface-marker confound on this corpus. *File: scatter_register_overlap.png*



AI Usage Statement

Generative AI tools were used during the development of this study to assist with code drafting, code debugging, figure generation, and language editing of manuscript text. All analytical decisions, lexicon application, statistical testing, interpretation of findings, and final verification of outputs were made by the author. All analyses reported in the manuscript were executed locally through the study scripts, inspected by the author, and cross-checked against the exported outputs. The author takes full responsibility for the accuracy, integrity, and interpretation of the work.

Data Availability

All response data, analysis scripts, and governance specifications are available at:

- CPFull pipeline: https://github.com/RayanBVasse/GenAI_Hermeneutics
- CPLite platform: https://github.com/RayanBVasse/Authors-Living-Literature
- Analysis scripts: `drift_comparison_analysis.py`, `register_overlap_analysis.py`
- Raw responses: `CPLite-Sonnet-15Qs/` and `CPFull-Sonnet-15Qs/`

References

Vasse, R. B. (2026). From Horizon to Heuristic: A governance fidelity instrument for detecting register drift in LLM-mediated reading. https://doi.org/10.5281/zenodo.19968329.

Vasse, R. B. (2026). Authorial Governance Reduces Register Driftin LLM Reading Companions: A matched-model across 10 canonical texts. https://doi.org/10.5281/zenodo.20133667