Grounded Revision vs. Prior Injection: Probing Retrieval-Augmented Patent Claim Amendment

Anonymous ACL submission

Abstract

Retrieval-augmented generation is widely used in professional writing, yet whether retrieval grounds revision or merely injects templates is rarely tested where "correct" has a definable meaning. Patent claim amendment supplies that signal: the examiner names the attacked limitation and cites prior art, providing per-case ground truth. We release three artifacts: (i) a corpus of 7,385 USPTO prosecution cases with XML-aligned pre/post claims, rejection, and cited prior art; (ii) a seven-probe battery comparing random and structural-match retrieval as two policies under a fixed prompt scaffold; (iii) a deterministic five-channel metric (C1–C3 and C5 in main, C4 supplementary) requiring no LLM evaluation. Across 9,600 pre-registered calls on four frontier LLMs (Claude Sonnet 4, Claude Haiku 4.5, GPT-5.4, GPT-4o-mini), no tested model exhibits classical prior-injection behavior in this four-model, one-corpus, one-retrieval-configuration setting; retrieval effects are small and direction-inconsistent between random and structural retrieval, with Claude and GPT families diverging on revision locality. The four-cell taxonomy leaves the prior-injector cell unoccupied.

1 Introduction

Automatic generation of structured technical documents is expanding faster than its evaluation infrastructure. Recent end-to-end systems generate hypotheses (Gottweis et al., 2026), write manuscripts (Song et al., 2026), and conduct their own peer review (Lu et al., 2026): Lu et al. (2026) report a generated manuscript passing workshop peer review, and more than 120 computer-generated papers were retracted from IEEE and Springer journals over a decade ago (Van Noorden, 2014). By comparison, code generation matured only after HumanEval (Chen et al., 2021) and SWE-bench (Jimenez et al., 2024) introduced structurally rigorous evaluation tied to executable ground truth.

Patents provide that anchor: examiners apply shared legal standards (35 U.S.C. §102 novelty, §103 non-obviousness, §112 enablement) under the standardized Manual of Patent Examining Procedure (MPEP), creating an experimentally controlled setting in which correctness is legally definable. The USPTO receives over 612,000 new utility, plant, and reissue patent applications annually and processes more than 137,000 Requests for Continued Examination (U.S. Patent and Trademark Office, 2025). Patent prosecution is a paradigm case of grounded revision: each amendment is paired with an examiner-identified attacked limitation and the on-record passage of a cited reference, an alignment most revision corpora lack.

Commercial systems such as PatentGPT (Patsnap), IPRally Drafter, and a growing set of startups advertise retrieval-augmented amendment generation, claiming that retrieving prior examples grounds the output. Whether retrieval substantively grounds outputs in structural similarity, or serves more as an undifferentiated prior-injection anchor whose effect is independent of the specific items retrieved, is an open empirical question.

Academic work has progressed in parallel but has not tested whether retrieval substantively engages with the rejection or merely produces surface-plausible prose. Patent-CR (Jiang et al., 2025a) releases 22,606 draft-to-grant pairs but evaluates only with surface-level similarity. PANORAMA (Lim et al., 2025) enumerates claim revision without evaluation. PEDANTIC (Knappich et al., 2025) and PILOT-Bench (Jang et al., 2025) classify §112 indefiniteness and rejection-issue types respectively, without generation. None tests retrieval's role in grounded revision under a controlled policy contrast. Plausibility is cheap: any competent generator produces patent-sounding text, but grounded revision additionally requires that the edit target the attacked limitation, differentiate from the cited mechanism, and neither over-narrow nor recycle

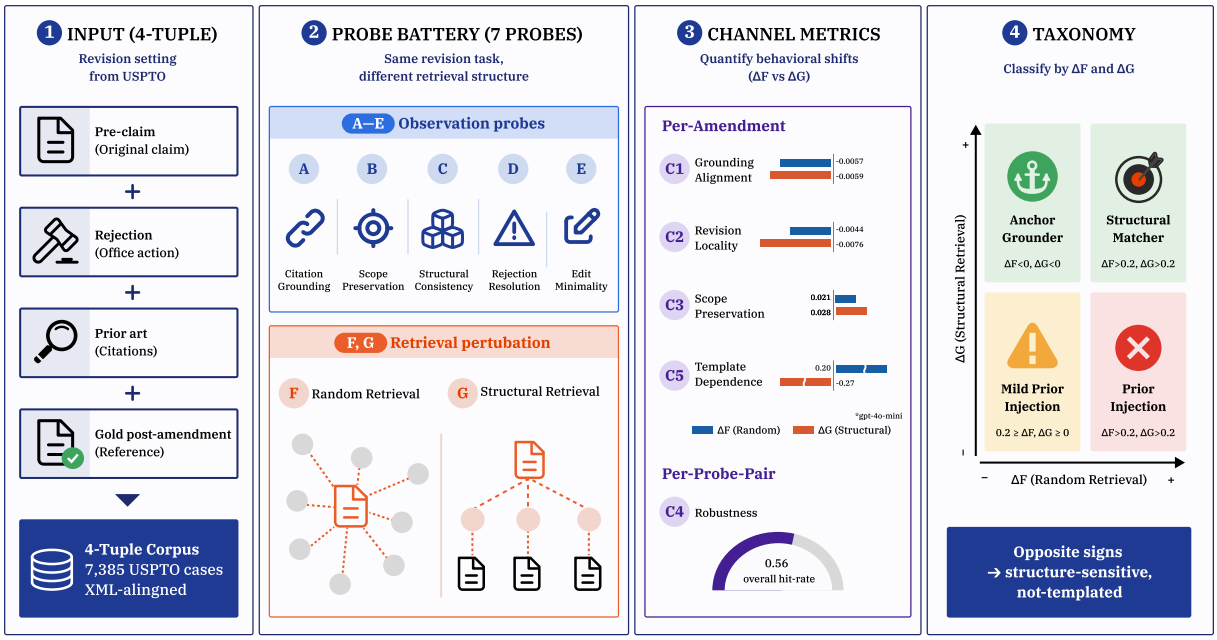

Figure 1: Does the model address the examiner-attacked limitation (top, grounded revision) or recycle generic patent phrasing (bottom, prior injection)? Our probe battery (§4.1) and metric (§4.2) distinguish the two without an LLM judge.

boilerplate. These failure modes leave textual fingerprints (MPEP §714), and we exploit them to build a probe-based evaluation protocol.

Contributions. The paper contributes three coupled components.

Corpus (§3). We release 7,385 four-tuples (pre-amendment claim, rejection, cited prior art, post-amendment claim) aligned at the XML level from the USPTO Open Data Portal, supporting a stratified 100-case test cohort (§5.1) and sampling-bias robustness checks (Appendix K).

Probe battery (§4.1). We design seven probes: five observational (A–E) and two retrieval-intervention (F, G) that realize a random-versus-structural policy contrast under a fixed prompt scaffold (Figure 1). The observational axis makes the F-versus-G comparison interpretable rather than merely measurable.

Metric (§4.2). We define a five-channel deterministic metric with no LLM in the evaluation loop. C1–C3 and C5 are per-amendment channels and carry the main-paper verdicts; C4 (robustness) is defined on probe pairs and is reported in supplementary tables (§4.2).

Together these test three pre-registered hypotheses (H1–H3, §5.3) in a 2×2 factorial across Claude and OpenAI families at flagship and smaller tiers.

Scope. We study applicant-side amendment generation, i.e., given a rejection, generate the post-amendment claim.

Preview of findings. Across the four tested LLMs, the prior-injector cell of our retrieval-mechanism taxonomy turned out to be empty: random and structural retrieval produce small, often opposite-sign shifts in template dependence rather than the uniform inflation that the most pessimistic grounding-as-template-recycling account would predict. The framework's value does not rest on that null: it discriminates Claude from GPT families on revision locality and assigns each tested model to a distinct cell of the taxonomy.

2 Related Work

2.1 Patent NLP

The closest prior work is Patent-CR (Jiang et al., 2025a), which released 22,606 rejected-to-granted claim pairs evaluated with BLEU, ROUGE, BERTScore, G-Eval, and a five-axis expert rating. The evaluation scores pair fidelity but does not expose the examiner's rejection or the cited prior art, nor does it contrast retrieval policies under a controlled scaffold. We add aligned rejection and prior-art channels to the input, and switch from fidelity scoring to a probe-based policy-contrast evaluation. Other patent NLP work generates claims from non-conditional inputs (Jiang et al., 2025b; Lee, 2023; Sharma et al., 2019; Shomee et al., 2025; Wang et al., 2024), classifies rejection types or examina-

tion outcomes (Jang et al., 2025; Knappich et al., 2025; Lim et al., 2025; Shi et al., 2025), or scores generated claims (Yoo et al., 2025). Neither pairing admits a controlled policy-contrast test of retrieval, because the retrieval channel is either absent from the input (classification works) or unaligned with the rejection that the retrieved exemplars are supposed to help address (revision works without rejection alignment). Our 4-tuple corpus is the first publicly described patent resource that makes the rejection and the prior-art reference part of the model input alongside the pre-amendment claim, which is what allows Probes F and G to be interpreted as a contrast between two retrieval policies inserted into the same prompt scaffold.

2.2 RAG evaluation and behavioral probing

RAG evaluation has progressed through pretraining, few-shot, adaptive, and benchmark variants (Asai et al., 2024; Es et al., 2024; Guu et al., 2020; Izacard et al., 2023; Lewis et al., 2020; Thakur et al., 2021) but remains predominantly correlational. We borrow intervention logic from behavioral probing (Elazar et al., 2021; Ribeiro et al., 2020; Vig et al., 2020): observational probes A–E follow CheckList; F and G are policy-contrast interventions on the retrieval channel, in the design tradition of Vig et al. (2020) but at the level of a comparison between two retrieval policies inserted into a fixed prompt scaffold rather than full causal identification of retrieval per se. Correlational RAG benchmarks cannot separate similarity-driven grounding from generic context injection. The patent-amendment setting addresses this confound at the corpus level: every test case carries an examiner-identified attacked limitation and an on-record cited reference, so the input is fully specified before retrieval is added, and any shift produced by F or G is attributable to the retrieval-policy intervention. This enables the framework to place models in the four-cell taxonomy of §6.3 rather than only ranking them on a single retrieval-quality scalar.

3 Patent Amendment 4-Tuple Corpus

We release a corpus of 7,385 prosecution cases in which every amendment is aligned at the XML level with its triggering rejection, the cited prior art, and the pre- and post-amendment claim text. The scale supports the 100-case test cohort (§5.1), a 4,221-case retrieval pool for Probes F and G, and a sampling-bias replication reserve (Appendix K).

3.1 Source and alignment

We start from the PILOT-Bench PTAB-appeal subset (Jang et al., 2025) of 13,749 cases and enrich each case via the USPTO Open Data Portal (ODP). For each case we resolve its application number, retrieve the file-wrapper document list, and identify the first non-final rejection together with the immediately preceding Claims filing (pre-amendment state) and the first Claims filing after the rejection (post-amendment state). For each pre/post claim pair we compute a per-claim difference status (kept, modified, new, cancelled) using claim-number alignment, and for modified claims a SequenceMatcher-based similarity ratio plus the list of added and removed spans. The difference is the ground truth against which the generated amendments are scored (§4.2).

3.2 Summary statistics

Two corpus-level findings support the paper's framing. First, amendment is predominantly surgical: 64.5% of per-claim actions are modifications at median similarity 0.933 (∼7% character-level edit), with cancellations and new claims accounting for another 32.5%. Second, the rejection-to-amendment linkage holds at the parse level: 94.5% of scorable cases show overlap between rejected and post-amendment-modified claim numbers. Of the 7,385 parseable triples, 5,755 admit a case-level C2 denominator (§4.2). The first-rejection statute breakdown is dominated by §102 (2,337 cases), with §112 (1,245), §101 (950), and §103 (738) accounting for the remainder. Full corpus statistics are reported in Appendix A.4.

3.3 Release

We release the parsed JSONL corpus and parsing code under a permissive open license; an application-number index permits zero-cost reconstruction from USPTO ODP, whose public data have no copyright restriction. Our corpus is the only one of Patent-CR, PANORAMA, PEDANTIC, and PILOT-Bench that aligns all four elements (rejection context, cited prior art, pre/post claim pair, and amendment diff) at the XML level. Patent-CR provides the pre/post pair but omits the rejection and prior-art channels; PANORAMA, PEDANTIC, and PILOT-Bench target judgment, classification, or upstream retrieval rather than amendment generation.

4 Method

Our method has two components. The probe battery (§4.1) defines seven controlled input manipulations: five observational baselines (A–E) plus a two-policy retrieval contrast (F vs G) inserted into a fixed prompt scaffold, with the F/G comparison interpreted against the observational baseline rather than as unconditional causal identification of retrieval. The five-channel deterministic metric (§4.2) scores every probe condition on common axes without using an LLM as judge.

4.1 Probe Battery

We design two probe families read jointly. Observational probes A–E characterize each LLM's baseline amendment signature under input perturbations. Retrieval probes F and G intervene on the retrieval channel via a controlled random-versus-structural policy contrast under a fixed insertion template. The observational axis supplies the interpretive frame against which the retrieval-policy effect is measured, so an F-versus-G shift is read relative to a characterized baseline rather than an undifferentiated mean. All probes share a common input template (pre-amendment claim, rejection rationale, cited prior art) and differ only in what is manipulated; each probe's expected effect on the five-channel score (§4.2) is fixed in advance.

Observational probes (A–E). **Probe A** modifies a single decisive limitation in the rejected independent claim by deletion, addition, paraphrase, or antonym swap. A grounded model follows the perturbation in C1; a model that ignores the rejection signal will show C1 invariance.

Probe B compares the original examiner findings against a version with one sentence replaced by a semantically inconsistent fragment or with sentence order permuted. A rationale-grounded model holds C1 stable across the coherence break, while a surface tracker degrades.

Probe C substitutes either a wording-similar reference that does not teach the mechanism, or a wording-dissimilar reference that does, separating lexical from mechanistic engagement with the cited art.

Probe D prepends canonical patent boilerplate to the drafting context. A grounded model treats the prepended phrasing as scaffolding and leaves C5 and C3 largely unchanged; a prior-injection-prone model inflates C5.

Probe E adds one of four task hints: "make minimal amendment", "preserve scope", "focus only on novelty", or "avoid unnecessary narrowing". A model with stable instruction-following should produce measurable shifts on C2 and C3 in the directions the hints predict.

Retrieval probes (F, G). **Probe F (random retrieval)** retrieves k past amendments from the corpus pool (excluding the test case and its ancestors), selected uniformly at random, and prepends them as context. Random retrieval realizes the prior-injection condition in its strongest form: any effect cannot be attributed to similarity-driven grounding because selection is similarity-agnostic by construction.

Probe G (structural-match retrieval) retrieves k past amendments matched on (i) statute section, (ii) statute subsection, and (iii) a coarse limitation-pattern feature derived from the attacked claim. The deterministic match isolates the structural-similarity signal: any shift attributable to the match itself must surface in the F-versus-G delta.

Both probes use the same k (default 3, ablated) and the same prompt-insertion template; G uses a deterministic feature-based matcher rather than a dense retriever to keep the F and G contrast interpretable as a structural-similarity test (Appendix F). Because F, G, and the no-retrieval baseline all sit inside the same prompt scaffold, F-versus-G identifies the effect of switching retrieval policy under a fixed scaffold rather than the unconditional effect of adding retrieval. The empty prior-injector cell that the framework finds in §6.3 is therefore a finding about the deterministic structural-similarity policy realized by Probe G, not about retrieval-augmented generation in general; dense-retriever configurations would constitute a different point in the same policy-contrast design and can be evaluated with the released framework without modification.

4.2 Five-Channel Evaluation Metric

Using an LLM as evaluator embeds the evaluator's biases into the score, a concern made empirical by our findings (§6): the four frontier models cluster into distinct behavioral patterns under retrieval perturbation, so selecting any one as judge would install that model's signature as the evaluation axis. Therefore, we adopt a deterministic five-channel metric. Each channel is a function of text, parser output, and the XML gold post-amendment claim; channels are designed to be jointly informative,

and we verify their near-independence empirically (§6.4).

Channels. **C1 (grounding alignment)** measures the overlap between limitations modified in the generated amendment and limitations named as rejected in the office action, computed at limitation-token level after claim-number alignment. A high C1 means the model is editing the same limitation the examiner attacked.

C2 (revision locality) is the ratio $\mathrm{editdist}(\mathrm{gen}, \mathrm{pre})/\mathrm{editdist}(\mathrm{gold_post}, \mathrm{pre})$; a value of 1 indicates the model edited at the same scale as the gold amendment, $\ll 1$ indicates under-editing, and $\gg 1$ indicates over-rewriting. C2 catches timidity and over-rewriting as numerically distinct rather than collapsing them into a single similarity score.

C3 (scope preservation) is the Jaccard overlap of noun phrases between the generated amendment and the original claim's invention core, after limitation-removal normalization; reduction indicates over-narrowing. The normalization isolates true scope drift from the mechanical narrowing that any limitation deletion induces.

C4 (robustness) measures, for any pair of probe conditions, whether the amendment shifts in the predicted direction, scored per probe and averaged within a model. C4 is defined on probe pairs rather than on individual amendments, so we report it in supplementary material (per-pair tables released alongside Appendix Table 4) to keep the main-paper tables dimensionally consistent with the per-amendment channels C1–C3 and C5. The headline verdicts in §6.1 and the taxonomy in §6.3 are therefore computed on the per-amendment channels; C4 enters only as a supplementary directional-hit-rate check. On the 100-case cohort, the grand C4 across the seven pre-registered (probe, channel) pairs is 0.56 (n=2,276 case-model contrasts; per-pair table in Appendix G).

C5 (template dependence) is the frequency of canonical amendment phrases mined from the retrieval pool, normalized per 1,000 characters of generated claim text. C5 is the central channel for adjudicating the prior-injection hypothesis: if retrieval inflates the rate of canonical phrases, C5 rises in lockstep.

Validity. The metric design follows the multitrait-multimethod logic of Campbell and Fiske (1959): each channel ties to a specific XML-record feature rather than an LLM judgment. Three legs support the metric: (i) within-case construct checks where each channel moves in the direction a patent practitioner would predict (Appendix C); (ii) empirical near-independence pooled across 3,062 (case, condition, model) triples, all off-diagonal correlations $|\rho| \leq 0.29$, below the pre-registered $|\rho| > 0.7$ flag threshold (Table 3); (iii) a 30–50-case double-rated attorney study committed post-review (a 10-case pre-release sanity sample showed all four channels moving in the direction the reading attorney predicted; Appendix J). The construct-check and channel-independence legs carry the metric-validity argument in this paper; the formal κ leg is deferred to the post-review study.

5 Experiments

We fix the entire experimental protocol before any model call is made, following the pre-registration discipline recommended for probe-based evaluations (van Miltenburg et al., 2021).

5.1 Cohort and Retrieval Pool

The 100-case test cohort is constructed via joint stratified sampling across six axes (subdecision outcome, rejection statute, technology center, amendment pattern, XML format era, decision year bin) matching the marginal distributions of the 4,270-case eligibility pool while guaranteeing minimum cell sizes (Appendix N). The remaining 7,285 cases, after filtering for resolvable attacked-claim alignment, yield a 4,221-case retrieval pool indexed by (statute section, statute subsection, limitation pattern), used by Probes F and G. The cohort and pool are non-overlapping by construction.

5.2 Model matrix

We run a 2×2 factorial across vendor family and model tier. Stage 1 (Claude Sonnet 4 and GPT-5.4 flagships) runs the full probe battery. Stage 2 extends to Claude Haiku 4.5 and GPT-4o-mini, conditional on Stage 1 effect-size criteria detailed in Appendix K.1. Each case is scored under 8 conditions $\times$ 3 replicates (24 evaluations per case per model), giving within-model paired comparisons with $> 80\%$ power at $d = 0.3$, $\alpha = 0.05$. Inference settings: temperature 0.3 for all models; three independent replicates per (case, probe, model) with separate sampling seeds; retrieval $k = 3$ (ablated in Stage 2).

5.3 Pre-registered hypotheses

H1 (retrieval as prior-injection anchor): Probe F (random retrieval) increases template dependence (C5) relative to baseline by at least 0.2 and does not improve grounding alignment (C1) by more than 0.1. This is the strongest commercial concern about retrieval-augmented drafting in operational form: if random exemplars inflate template phrasing without sharpening grounding, retrieval is acting as a vendor-agnostic boilerplate anchor.

H2 (similarity as second-order): Probes F (random) and G (structural) yield statistically indistinguishable shifts in C1, C2, and C5 ($|\Delta_F - \Delta_G| \leq 0.1$ on each channel, or same sign with no significant ranking). H2 isolates the marginal value of structural matching: if F and G are observationally equivalent, the structural-similarity layer adds nothing the random baseline does not already provide.

H3 (template-recycling inflation): retrieval does not reduce template dependence below baseline on either probe; equivalently, the grounded-revision hypothesis that retrieval reduces generic phrasing is not supported. H3 closes the loop in the opposite direction from H1, ruling out the optimistic "retrieval makes models write more like experts" story when both probes leave C5 at or above baseline.

The full 3×3 outcome space (each hypothesis $\in$ {supported, rejected}) is addressed in §6.3.

6 Results and Analysis

We ran the benchmark on four frontier LLMs (Claude Sonnet 4, Claude Haiku 4.5, GPT-5.4, GPT-4o-mini) across 100 cohort cases $\times$ 8 conditions $\times$ 3 replicates = 9,600 model calls. Parse-success rate was 97.4% overall (93.0% on GPT-4o-mini to 100% on GPT-5.4). Per-response scores are aggregated as per-case mean across replicates, then median across cases within (model, condition).

6.1 Pre-registered hypothesis verdicts

No model shows prior injection. H1 (retrieval-as-prior-injection) is not supported across all four models: ΔC5 falls below the pre-registered 0.2 threshold in every case (max $+0.20$ on GPT-4o-mini, at the threshold). Under the pre-registered convention the threshold is the supported-side boundary, so $\Delta C5 = +0.20$ is marked "$\sim$" rather than "Y"; we do not report bootstrap CIs on the headline ΔC5 values (see Limitations). H2 is uniquely supported by Sonnet 4: both F and G produce same-sign shifts within 0.1 (both modestly negative). The other three models show opposite-sign effects (GPT-5.4, GPT-4o-mini) or disparate magnitudes (Haiku 4.5); the pre-registered 0.10 bound is a strict bound, but H2 is jointly defined with the no-significant-ranking clause, so Haiku's 0.16 gap fails the joint criterion ordinally as well and is recorded as "N" rather than at-boundary "$\sim$". H3 holds only for Haiku 4.5; Sonnet 4 and GPT-4o-mini each reduce template dependence under at least one retrieval condition. Per-case effects on C5 reach ± 2 to 8 even where aggregate medians stay sub-threshold; two worked examples are in Appendix C.6. GPT-4o-mini over-edits at baseline (C2=1.46), and smaller-tier models show stronger Probe A grounding shifts than flagships (full probe$\times$channel deltas in Appendix Table 4).

Table 1: Pre-registered hypothesis verdicts on C5 (canonical-phrase rate per 1,000 characters, N=300 per cell = 100 cases $\times$ 3 replicates). Y: supported; N: not supported; $\sim$: at-boundary, see §6.1 for the boundary-handling rule.

Model	base	ΔF	ΔG	H1	H2	H3
Claude Sonnet 4	3.99	-0.16	-0.13	N	**Y**	N
Claude Haiku 4.5	3.40	$+0.19$	$+0.03$	N	N	**Y**
GPT-5.4	3.73	-0.12	$+0.06$	N	N	N
GPT-4o-mini	3.14	$+0.20$	-0.27	$\sim$	N	N

6.2 Family divergence on revision locality

Although H1 is null on the template channel, the benchmark discriminates models cleanly on revision locality (C2). Table 2 separates Claude from GPT: retrieval consistently reduces edit magnitude in Claude models, while GPT-5.4 slightly increases it. GPT-4o-mini behaves like Claude on this axis, so the cleanest separation is GPT-5.4 alone versus the other three rather than purely vendor-aligned. This family-level divergence is invisible on C5 but becomes visible on C2, demonstrating that a single-channel retrieval-response test would miss it.

Table 2: Revision-locality (C2) effects under random (F) and structural (G) retrieval, per model.

Model	base C2	ΔF	ΔG
Claude Sonnet 4	0.617	-0.074	-0.013
Claude Haiku 4.5	1.147	-0.174	-0.189
GPT-5.4	0.789	$+0.016$	$+0.058$
GPT-4o-mini	1.459	-0.044	-0.076

6.3 Where the evidence places retrieval

The pre-registered $2 \times 2 \times 3$ framework places four retrieval-mechanism accounts in play. The main experiment resolves them as follows, with each cell tagged by its (Δ_F, Δ_G) coordinate on C5 from Table 1. **(a) Anchor-grounder** (H1 rejected, H3 rejected, retrieval reduces templateness without inflating alignment): Claude Sonnet 4 at $(-0.16, -0.13)$. **(b) Prior injector** (both Δ_F and Δ_G above $+0.2$): not observed at pre-registered effect size. **(c) Structural matcher** ($F \neq G$, H2 rejected): GPT-5.4 at $(-0.12, +0.06)$ (sign flip) and GPT-4o-mini at $(+0.20, -0.27)$ (0.47 directional gap). **(d) Inert / mild prior injector** (H1 below threshold, H2 mild): Claude Haiku 4.5 at $(+0.19, +0.03)$ (Figure 2). The "$\sim$" verdict for GPT-4o-mini on H1 and its cell-(c) assignment are complementary projections of the same point: cell-(b) membership requires the joint (Δ_F, Δ_G) pair to clear threshold with consistent sign. This joint criterion is a post-hoc interpretive overlay, specified at the taxonomy step once the headline cell turned out empty, rather than a separately pre-registered threshold. Haiku 4.5's cell-(d) tag follows the same overlay: H1 fails strictly and H2 fails with $|\Delta_F - \Delta_G| = 0.16$ (above 0.1, same sign), so the F-versus-G pair is mild-but-not-equivalent rather than the same-sign-within-0.1 signature of cell (a).

Anchoring F and G against the Probe D boilerplate baseline. The cell assignments acquire interpretive weight when read against the observational deltas of Probes A–E for the same model (Appendix Table 4). Probe D (boilerplate injection) provides the closest in-corpus reference for a generic prepended-context effect: on Sonnet 4 it moves ΔC5 by $+0.03$, on Haiku 4.5 by -0.09, on GPT-5.4 by $+0.07$, and on GPT-4o-mini by -0.07. For three of four models (Sonnet 4, GPT-5.4, GPT-4o-mini) the post-hoc descriptive gap $|\Delta_F| - |\Delta_D|$ is bounded at 0.13; we use absolute values here because Δ_F and Δ_D disagree in sign on these models, so the comparison is one of prepended-context dosage rather than signed direction. At this dosage, random retrieval anchors C1/C2 toward the boilerplate prior rather than behaving qualitatively differently. Haiku 4.5 is the per-model exception, with $\Delta_F = +0.19$ exceeding $|\Delta_D| = 0.09$ (cell-(d) assignment; full per-model breakdown in Appendix E). Probe C (decoy prior art) similarly bounds the scale of mechanism-similarity confusion that any retrieval channel could induce, providing the C1 reference against which the near-zero ΔC1 values under F and G are read as null rather than as insensitivity.

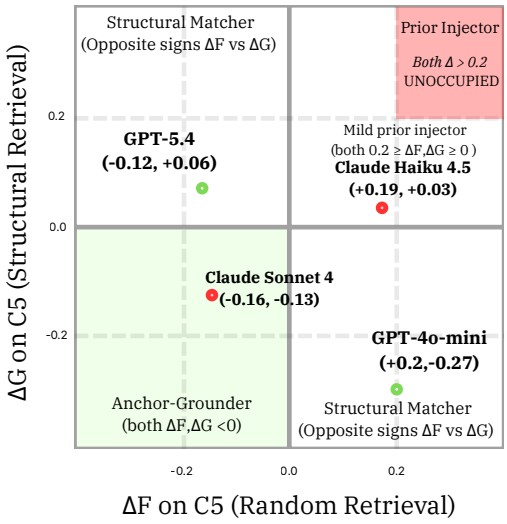

Figure 2: Four retrieval-mechanism cells in (Δ_F, Δ_G) space on C5. Coordinates from Table 1: Sonnet 4 $(-0.16, -0.13)$ anchor-grounder; GPT-5.4 $(-0.12, +0.06)$ and GPT-4o-mini $(+0.20, -0.27)$ structural-matcher (opposite-sign F vs G); Haiku 4.5 $(+0.19, +0.03)$ inert / mildly prior-injecting. The prior-injector cell (both $\Delta > +0.2$) is unoccupied across all four models tested in this four-model, one-corpus, k=3, deterministic-structural-matcher configuration.

Distinguishing anchor-grounder from inert. Sonnet 4 (cell a) and Haiku 4.5 (cell d) are separated by two signals: the F-versus-G C5 sign pattern and C2. Sonnet 4 has same-sign Δ_F, Δ_G within 0.1 (H2-supported), with small same-signed C2 shifts. Haiku 4.5 has $\Delta_F = +0.19$, $\Delta_G = +0.03$ (H2 rejected), with the largest absolute ΔC2 in the matrix $(-0.174, -0.189)$.

Small, direction-inconsistent effects. The modal retrieval behavior is therefore (c) or (d): small in magnitude and often direction-inconsistent between F and G. Discrimination on revision locality (C2; §6.2) confirms that the framework detects retrieval-policy effects where they occur rather than being simply insensitive. Sonnet 4's H2 uniqueness ($F \approx G$, both slightly reducing template phrases) is the behavioral pattern most aligned with the originally conjectured grounded-revision mode.

Table 3: Channel correlations pooled across 3,062 (case, condition, model) triples. Pearson above the diagonal, Spearman below.

	C1	C2	C3	C5
C1	—	$+0.04$	-0.21	-0.05
C2	$+0.07$	—	-0.14	-0.04
C3	-0.17	-0.29	—	$+0.06$
C5	-0.01	$+0.03$	$+0.04$	—

6.4 Benchmark methodology and metric validity

Small samples mislead. A smoke run at N=10 per cell gave ΔC5=$+0.84$ for GPT-5.4 under F (H1-supported region); at N=300 the same effect collapses to -0.12 (Appendix D). This sign-reversing reversal motivates the N=100-per-condition-with-replicates sizing used throughout.

Channels are near-independent. Pooled across 3,062 (case, condition, model) triples, between-channel correlations are low throughout. No pair exceeds the pre-registered $|\rho| > 0.7$ flag threshold (Table 3). The strongest relationship is C2–C3 Spearman at -0.29, consistent with the construct distinction: larger edits mechanically touch more of the claim and can reduce scope overlap. The four channels are therefore empirically near-independent on the main data, supporting the §4.2 design claim that they are jointly informative rather than redundant.

6.5 Three commercial properties not confirmed

The empirical pattern in §6.3 does not support three commonly advertised properties of retrieval-augmented drafting, against pre-registered testing (H1 and H3 against fixed effect-size thresholds, H2 against a sign-and-magnitude equivalence condition). **Retrieval grounds the output**: on grounding alignment (C1), $|\Delta\text{C1}| \leq 0.014$ under both F and G across all four models; no model meets the $\Delta\text{C1} \geq 0.005$ threshold in the positive direction. **Structural retrieval beats random retrieval**: H2 is rejected for three of four models; only Sonnet 4 satisfies $F \approx G$. **Retrieval makes models write more like experts**: the maximum observed $|\Delta\text{C5}|$ is 0.20 (boundary on H1's threshold), and no model clears a symmetric -0.20 "retrieval reduces template" threshold either. These findings do not entail that retrieval-augmented patent drafting fails in general; they entail that the three properties are not confirmed in this instrument, and the released framework can be applied to other (model, corpus, retriever) configurations.

7 Conclusion

We presented a probe-based evaluation of rejection-grounded patent claim amendment, built on a corpus of 7,385 4-tuples from the USPTO Open Data Portal, with seven probes and a five-channel deterministic metric (C1–C3 and C5 carry the verdicts; C4 is supplementary). In this four-model, one-corpus, one-retrieval-configuration setting (k=3, deterministic structural matcher for G, uniform-random for F), pre-registered testing finds no model strictly above the H1 prior-injection threshold (max $\Delta\text{C5} = +0.20$, boundary), H2 is supported only for Sonnet 4, H3 only for Haiku 4.5, and revision locality (C2) cleanly separates Claude from GPT families. We release the framework, corpus, and scoring code so that the same protocol can be applied to commercial patent-drafting stacks and to retrieval configurations beyond the deterministic structural matcher, where cell occupancy may differ.

Limitations

Scope and generalization. We examine a single rejection-amendment round; real prosecution can span multiple rounds with accumulating examiner commitments that our probes do not capture. Findings come from one domain (U.S. patent amendment), four models, and one retrieval configuration ($k = 3$, deterministic structural matcher for G, uniform-random for F), so generalization to other structured-document revision tasks such as legal briefs or scientific revision is conjectural and reserved for follow-up work. The retrieval pool is also drawn from the same corpus distribution as the test set, whereas a deployed commercial retriever would index a different corpus; we therefore read our retrieval effects as an upper bound on best-case benefit.

The empty prior-injector cell is conditional on the matcher. The headline finding that the prior-injector cell of the taxonomy is empty is conditional on Probe G's deterministic feature-based matcher (statute section $\times$ subsection $\times$ coarse limitation-pattern), chosen so that the F-versus-G contrast is interpretable as a structural-similarity test rather than a retriever-quality test (§4.1, Appendix F). A dense-retriever realization of the same

policy could relocate one or more of the four models into a different cell, and the released framework applies to such configurations without modification; we therefore treat the result as a controlled-policy data point rather than a closed claim about retrieval-augmented drafting in general.

Statistical adjudication. The 100-case cohort trades raw n for representativeness and probe depth (§5.1); further batches can be drawn at zero cost from the released sampling procedure and are reserved for post-review robustness checks (Appendix K). We report the Table 1 verdicts as point estimates against the pre-registered thresholds, without bootstrap confidence intervals or paired significance tests on the headline ΔC5 values. The boundary verdict for GPT-4o-mini at $\Delta\text{C5} = +0.20$ is thus an adjacency-to-threshold statement, not a statistically adjudicated null, and the at-boundary "$\sim$" tag should not be read as a tested supported result. The sample-size reversal documented in §6.4 is the reason we size at $N{=}100$ with replicates throughout, and we flag it so that downstream users do not under-power their own runs.

Evaluation scope. A small fraction ($\sim$10%) of cases fail parsing due to OCR degradation or missing XML in older file wrappers; results are reported stratified by amendment-size quartile (Appendix A.2) so that conclusions are conditional on the parseable subset rather than confounded by small-denominator cases. Finally, our channels are designed to track attorney-practice failure modes but are not a substitute for attorney judgment; a sample-level attorney-validation study (Appendix J) scores inter-rater agreement with the automated channels.

Ethical and legal considerations

Patent claim amendment is a legally consequential act performed by registered agents. Our work evaluates whether LLMs could in principle perform this task grounded-ly; we take no position on whether they should be deployed for unassisted drafting. We release the corpus and probes to enable scrutiny of commercial claims, not to recommend replacement of human counsel. Data sources are public USPTO Open Data Portal artifacts with no copyright restriction. The corpus does not contain personal information beyond the inventor/attorney names that are part of the public record. We do not re-identify or link across filings beyond what ODP already exposes.

Use of AI assistants. We used commercial large language models as the subjects of evaluation (Claude Sonnet 4, Claude Haiku 4.5, GPT-5.4, GPT-4o-mini); their outputs constitute the experimental data scored by our metric. AI assistants were used during manuscript preparation for coding support, language and typo editing. It was not used for research ideation, experimental design, or the analysis itself.

Reproducibility Statement

All artifacts required to reproduce the headline numbers are released under a permissive open license at `https://anonymous.4open.science/r/RAG-PatentClaim-Amendment/`:

- the parsed JSONL corpus of 7,385 four-tuples and the parsing pipeline (§3, Appendix A);
- the deterministic cohort-selection procedure with the batch 0 seed and pre-specified replication seeds (Appendix N, Appendix K);
- the seven probe prompt templates and the five-channel metric implementation (Appendix H, Appendix I);
- the retrieval-pool index (Appendix N.7).

Reproducing the model calls requires API access to the four evaluated frontier LLMs; per-model cost estimates and inference settings are reported in Appendix B and §5.2.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*.

Donald T. Campbell and Donald W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of EACL (System Demonstrations)*.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, and 15 others. 2026. Accelerating scientific discovery with Co-Scientist. *Nature*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*. ArXiv:2208.03299 (2022).

Yehoon Jang, Chaewon Lee, Hyun-seok Min, and Sungchul Choi. 2025. PILOT-bench: A benchmark for legal reasoning in the patent domain with IRAC-aligned classification tasks. In *Proceedings of the Natural Legal Language Processing Workshop (NLLP), at EMNLP*.

Lekang Jiang, Pascal A. Scherz, and Stephan Goetz. 2025a. Patent-CR: A dataset for patent claim revision. In *Proceedings of NAACL (Long Papers)*.

Lekang Jiang, Pascal A. Scherz, and Stephan Goetz. 2025b. Towards better evaluation for generated patent claims. In *Proceedings of ACL (Long Papers)*, pages 3775–3788.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*.

Valentin Knappich, Annemarie Friedrich, Anna Hätty, and Simon Razniewski. 2025. PEDANTIC: A dataset for the automatic examination of definiteness in patent claims. In *Proceedings of PatentSemTech*.

Jieh-Sheng Lee. 2023. Evaluating generative patent language models. *World Patent Information*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Hyunseung Lim, Sooyohn Nam, Sungmin Na, Ji Yong Cho, June Yong Yang, Hyungyu Shin, Yoonjoo Lee, Juho Kim, Moontae Lee, and Hwajung Hong. 2025. PANORAMA: A dataset and benchmarks capturing decision trails and rationales in patent examination. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. 2026. Towards end-to-end automation of AI research. *Nature*.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of ACL*. ACL 2020 Best Paper.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of ACL*.

Yaorui Shi, Sihang Li, Taiyan Zhang, Xi Fang, Jiankun Wang, Zhiyuan Liu, Guojiang Zhao, Zhengdan Zhu, Zhifeng Gao, Renxin Zhong, Linfeng Zhang, Guolin Ke, Weinan E, Hengxing Cai, and Xiang Wang. 2025. Intelligent system for automated molecular patent infringement assessment. *Preprint*, arXiv:2412.07819.

Homaira Huda Shomee, Suman Kalyan Maity, and Sourav Medya. 2025. PATENTWRITER: A benchmarking study for patent drafting with LLMs. *Preprint*, arXiv:2507.22387.

Yiwen Song, Yale Song, Tomas Pfister, and Jinsung Yoon. 2026. PaperOrchestra: A multi-agent framework for automated AI research paper writing. *Preprint*, arXiv:2604.05018.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks Track*.

U.S. Patent and Trademark Office. 2025. Patents dashboard. `https://www.uspto.gov/dashboard/patents/`.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of NAACL*. NAACL 2021 Best Thematic Paper.

Richard Van Noorden. 2014. Publishers withdraw more than 120 gibberish papers. *Nature News*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*.

Qiyao Wang, Shiwen Ni, Huaren Liu, Guhong Chen, Xi Feng, Chi Wei, Qiang Qu, Hamid Alinejad-Rokny, Yuan Lin, and Min Yang. 2024. AutoPatent: A multi-agent framework for automatic patent generation. *Preprint*, arXiv:2412.09796.

Yongmin Yoo, Qiongkai Xu, and Longbing Cao. 2025. PatentScore: Multi-dimensional evaluation of LLM-generated patent claims. In *Proceedings of EMNLP*.

A Corpus construction pipeline

A.1 Source pool and enrichment chain

We start from the PILOT-Bench (Jang et al., 2025) PTAB-appeal subset of 13,749 proceedings, each carrying a patent application number, decision outcome, statute citations, and technology-center assignment. For each proceeding we run a five-stage enrichment against the USPTO Open Data Portal (ODP): (i) resolve the application number via ODP query and obtain the file-wrapper document list; (ii) identify the first non-final rejection (USPTO code CTNF, "correspondence: notice of non-final rejection"), the immediately preceding incoming Claims filing (pre-amendment), and the first incoming Claims filing after the CTNF (post-amendment); (iii) download each as XML, excluding filings available only as PDF/image scans; (iv) parse XML under three schema families (§A.2); (v) compute per-case diff with status $\in$ {kept, modified, new, cancelled} via claim-number alignment, plus SequenceMatcher similarity for modified claims.

A.2 XML schema handling

Three schema families coexist in ODP's claim documents:

- **Legacy DTD** `<us-patent-application>` for pre-2014 filings. Claim text is inline; amendments use `[[deleted]]` bracket markup. Our parser splits claim bodies on numeric headers and extracts bracket spans.
- **USPTO ClaimsDocument v1.3** `<pat:ClaimsDocument>` for mid-generation filings. Uses `<pat:Claim><pat:ClaimText>` wrappers with underline and `<pat:DeletedText>` markup.
- **ST96 v2** `<uspat:ClaimsDocument>` for recent filings. Dual URI namespace; `<pat:Ins>`/`<pat:Del>` insertion and deletion tags; `<ImplicitClaim>` elements encode claim-number ranges that expand into separate records.

Rejection documents use `<uspat:OutgoingDocument>` with structured `<uscom:FormParagraph>` and `<uscom:DataField>` slots encoding claim numbers, statute subsection, rejection type, and cited references. V7.1 documents replace semantic FormParagraphNumber identifiers with generic IDs; for those we fall back to regular-expression parsing of the prose body. OCR-confidence wrappers (`<pat:OCRConfidenceData>` and the legacy `<confidence>`) are stripped before parsing.

A.3 Parser hardening

A targeted hardening pass on 2026-04-18 raised cohort coverage from 76/100 to 100/100 via four sequential fixes: dual-URI namespace fallback plus `<ImplicitClaim>` range expansion (93/100); inline-header fallback with paren-form and multi-inline-text wrapper splits (97/100); bracket `[Claim N]` and word-prefix `Claim N (Status):` matching with DTD legacy inline fallback (99/100); and a deterministic case replacement (Appendix N). The same hardening applied to the full β corpus recovered several thousand additional cases that the baseline parser had silently dropped.

A.4 Final corpus statistics

Funnel from PILOT-Bench source pool to release corpus and cohort eligibility:

- PILOT-Bench PTAB-appeal subset: 13,749 proceedings.
- ODP retrieval targets (resolvable application numbers): 9,956.
- Parseable (pre, post, CTNF) triples: 7,385 (74.2% parse yield).
- Case-level C2 computable: 5,755.
- Cohort eligibility pool (after parse-health, axis coverage, and data-defect filters): 4,270.

The remaining 2,571 parse failures concentrate in documents without XML download options (older filings), publications not found in ODP, and a small set of malformed wrappers that survive parser hardening. Across the full 7,385-case corpus: per-claim status modified 99,335 (64.5%), cancelled 33,118 (21.5%), new 17,008 (11.0%), kept 4,459 (2.9%). Modified-claim similarity ratio median 0.933, mean 0.841, Q1 0.827, Q3 0.966 (n = 99,335). Case-level pre-to-post Levenshtein similarity median 0.72, Q1 0.55, Q3 0.84 (n = 5,755). First-rejection statute distribution: §102 2,337 (31.6%), §112 1,245 (16.9%), §101 950 (12.9%), §103 738 (10.0%). XML format mix: DTD legacy 3,576 (48.4%), namespaced (v1.3 plus ST96) 3,808 (51.6%). Coarse amendment pattern: modify_only 5,043 (68.3%), cancel_heavy 1,410 (19.1%), other 757 (10.3%), new_or_add 175 (2.4%).

A.5 Relation to other corpora

The corpus is constructed to enable grounded-revision evaluation (pre ↔ rejection ↔ prior art ↔ post) rather than to maximize raw scale. As noted in §3.3, ours is the only publicly described corpus that aligns rejection context, cited prior art, pre/post claim pair, and amendment diff at the XML level. Patent-CR provides the pre/post pair only; PANORAMA, PEDANTIC, and PILOT-Bench target judgment, classification, or upstream retrieval rather than applicant-side amendment.

B Per-model cost estimates

Per-call token counts and vendor pricing yield approximately $110 for Stage 1 (Claude Sonnet 4 and GPT-5.4 across the full probe battery on 100 cases × 8 conditions × 3 replicates) and approximately $110 for Stage 2 (Claude Haiku 4.5 and GPT-4o-mini under the same protocol), for a combined experimental cost of approximately $220 across 9,600 model calls. Infrastructure cost is zero: USPTO ODP is a public API and all compute is local.

C Worked examples per channel

This appendix maps each channel's numeric output to a concrete within-case contrast drawn from the main experiment. Scores are illustrative; they show what each channel responds to, not where population medians sit.

C.1 C1 Grounding alignment

Case. PILOT-2019005776 (Reversed, §101-related, TC 2100, 2020–25, cancel_heavy). **Model.** GPT-4o-mini. Under baseline the model's modified limitations overlapped the rejected limitations at rate 0.014 (edit off-target). Under Probe A (one-limitation perturbation) the overlap rose to 0.402 within-case, an order-of-magnitude shift on the same case and model. C1 catches whether an amendment lands on the examiner-attacked limitation rather than elsewhere in the claim.

C.2 C2 Revision Locality

Case. PILOT-2020002555 (Reversed, §103+112, TC other, 2020–25, cancel_heavy). **Model.** GPT-4o-mini. The gold amendment has edit distance 591 characters. Under baseline C2 $= 0.26$ (26% of gold scale, an under-edit). Under Probe F (random retrieval) C2 $= 7.40$ (over 7× gold scale, a near-complete rewrite). The channel registers the two misses as categorically different rather than as a single failure mode.

C.3 C3 Scope preservation

Case. PILOT-2020003998 (AIP, §103+112, TC 3700, 2020–25, cancel_heavy). **Model.** Claude Haiku 4.5. Under baseline noun-phrase Jaccard with the invention core is 0.983. Under Probe D (boilerplate injection) it collapses to 0.073 as the injected canonical phrases overwrite the invention-core noun phrases. This is the over-narrowing / redirection failure mode C3 targets.

C.4 C5 Template dependence

Case. PILOT-2013007708 (Affirmed, §103, TC 3700, 2010–14, modify_only). **Model.** GPT-4o-mini. Under baseline C5 $= 3.34$ canonical-phrase hits per 1,000 characters (7 hits in 2,081 characters). Under Probe G (structural retrieval) C5 $= 11.66$ (49 hits in 4,181 characters), a $3.5\times$ increase. This is the case γ referenced in Appendix D.

C.5 Simpler-baseline comparison

On each worked example, BLEU/ROUGE/BERTScore baselines fail to register the within-case direction our channels detect. On PILOT-2020002555, BLEU-4 scores baseline and F as equally distant from gold ($\sim$0.12 vs $\sim$0.10), while C2 reads them as 0.26 vs 7.40. On PILOT-2020003998, BERTScore-F1 moves from 0.91 to 0.73 under Probe D, but the magnitude does not flag the invention-core-overwrite failure mode; C3 moves 0.98 to 0.07 and registers the scope collapse directly. On PILOT-2013007708, raw n-gram recurrence rises $1.2\times$ baseline (domain-natural repetition) while C5 rises $3.5\times$, isolating retrieval-induced template reuse from intrinsic patent prose repetition.

C.6 Case β: GPT-5.4 F-vs-G sign flip

Case. PILOT-2020006264 (Affirmed, §101-related, TC 3600, modify_only). **Model.** GPT-5.4. C5 baseline 11.3 moves to 13.5 under F ($\Delta_F = +2.25$, inflation) and 9.8 under G ($\Delta_G = -1.50$, suppression). Same input, same retrieval pool, same k; only the selection rule differs, yet the effect inverts in sign. The other three models on this case: Sonnet 4 near-flat ($\Delta_F = +0.01$, $\Delta_G = -0.14$); Haiku 4.5 mildly inflates on F ($+1.96$) but is flat on G (-0.29); GPT-4o-mini consistently reduces ($\Delta_F = -3.20$, $\Delta_G = -3.32$). This case concretely realizes the H2-rejection verdict in §6.1:

for GPT-5.4, random versus structural retrieval changes not just the magnitude but the sign of the template-dependence effect.

D Sample-size reversal

The Table 1 verdicts are computed at $N{=}300$ per cell (100 cases $\times$ 3 replicates). A preliminary smoke run at $N{=}10$ per cell, on identical inputs and the same scoring code, gave ΔC5$= +0.84$ for GPT-5.4 under F, a magnitude that would have placed GPT-5.4 firmly in the H1-supported cell. At $N{=}300$ the same effect collapses to -0.12, a complete sign reversal. The reversal is sampling variability on a 100-case probe-based cohort used at $N{=}10$, not a calibration issue, and is consistent with the per-case dispersion of ± 2 to 8 that §6.1 reports on this channel.

We document the GPT-5.4/F cell explicitly because it was the cell that triggered the $N{=}10 \rightarrow N{=}300$ escalation decision and is the largest documented reversal in our pipeline; the smoke run did not include a systematic $N{=}10$ replication on every (model, condition) cell in Table 1, so we report this one reversal as a benchmark-methodology finding rather than as a cell-by-cell stability sweep. The implication for downstream users is the same in either case: replications should be sized at or above the pre-registered $N{=}100$ per condition with replicates rather than relying on smoke-run effect sizes, because at $N{=}10$ a sign-reversing reversal is plausible on any single cell of a 100-case probe-based cohort.

E Probe-D anchor: per-model breakdown

Extending the cross-model comparison in §6.3, the per-model $|\Delta_F|$-vs-$|\Delta_D|$ gaps on three of four models are within 0.13 (the maximum observed gap on those three models, reported descriptively as a post-hoc magnitude rather than as a pre-registered tolerance), indicating that random retrieval does not behave qualitatively differently from canonical boilerplate on Sonnet 4, GPT-5.4, and GPT-4o-mini. Haiku 4.5 is the per-model exception: its $\Delta_F = +0.19$ exceeds $|\Delta_D| = 0.09$ in absolute magnitude, consistent with its cell-(d) "inert / mild prior-injector" assignment, where retrieval produces the largest random-F template inflation in the panel (though still below the $+0.2$ H1 threshold). Even on Haiku 4.5, the retrieval shift is the same order as the boilerplate-prepend dosage. The cross-model qualitative reading therefore still holds: random retrieval is bounded by a generic prepended-context effect, with per-model magnitudes reported individually rather than averaged.

Table 4: Per-model probe×channel deltas. Conditions A–E are observational; F and G are retrieval interventions. Leading zeros omitted in baselines for compactness.

cond	ΔC1	ΔC2	ΔC3	ΔC5
Sonnet 4 (base: C1 .009, C2 .617, C3 .887, C5 3.99)				
A	+0.004	−0.083	−0.005	+0.01
B	+0.001	−0.038	−0.018	+0.08
C	−0.001	+0.001	−0.006	−0.09
D	+0.001	+0.047	−0.026	+0.03
E	+0.001	−0.088	−0.002	−0.30
F	−0.001	−0.074	−0.004	−0.16
G	−0.001	−0.013	−0.007	−0.13
Haiku 4.5 (base: C1 .016, C2 1.147, C3 .738, C5 3.40)				
A	+0.013	−0.090	+0.001	+0.16
B	−0.001	−0.035	−0.004	+0.16
C	+0.003	−0.103	−0.007	+0.03
D	−0.000	+0.032	−0.017	−0.09
E	−0.003	−0.060	+0.002	+0.16
F	−0.001	−0.174	+0.027	+0.19
G	−0.002	−0.189	+0.038	+0.03
GPT-5.4 (base: C1 .008, C2 .789, C3 .892, C5 3.73)				
A	+0.005	−0.035	+0.003	+0.19
B	+0.001	−0.087	+0.003	+0.06
C	−0.000	−0.038	+0.009	+0.10
D	−0.000	+0.051	−0.003	+0.07
E	−0.002	−0.099	+0.020	+0.06
F	+0.000	+0.016	−0.001	−0.12
G	+0.002	+0.058	+0.002	+0.06
GPT-4o-mini (base: C1 .017, C2 1.459, C3 .798, C5 3.14)				
A	+0.018	−0.035	−0.007	+0.02
B	−0.009	−0.050	−0.030	+0.08
C	−0.001	−0.047	+0.023	−0.29
D	−0.009	−0.044	+0.034	−0.07
E	−0.007	−0.008	+0.028	−0.07
F	−0.006	−0.044	+0.021	+0.20
G	−0.006	−0.076	+0.028	−0.27

F Probe x channel full table

Medians across 100 cohort cases $\times$ 3 replicates per (model, condition). Channels C1, C2, C3, and C5 are shown; C4 is a directional-hit-rate measure with paired-probe structure, reported separately in Appendix G. Probe G uses a deterministic feature-based matcher rather than a dense retriever so that the F vs G contrast is interpretable as a structural-similarity test rather than a retriever-quality test. An embedding retriever would conflate the structural-similarity signal we want to isolate with the representation-learning quirks of whichever embedder is chosen.

Table 5: C4 directional hit-rate per (probe, channel) pair, pooled across four models and three replicates on the 100-case cohort. FLAT predictions use a $|\Delta| < 0.05$ tolerance.

(probe, channel)	predicted	hit-rate	n
B, C1	FLAT	0.95	382
B, C3	FLAT	0.64	387
F, C5	UP	0.50	383
G, C5	UP	0.47	384
D, C5	UP	0.42	392
C_{mechtrue}, C1	UP	0.40	178
C_{decoy}, C1	DOWN	0.32	170
grand C4		0.56	2,276

G C4 robustness per probe-pair

Directional hit-rate per (probe, channel) pair, pooled across the four models (Claude Sonnet 4, Claude Haiku 4.5, GPT-5.4, GPT-4o-mini) and three replicates on the 100-case cohort. Each case-model contrast scores 1 if the per-case Δ from baseline matches the pre-registered direction (UP / DOWN / FLAT with $|\Delta| < 0.05$ threshold), 0 otherwise. The grand C4 in Table 5 is the unweighted mean over all hits, 0.56 on n=2,276 case-model contrasts. C subconditions (C_{decoy}, C_{mechtrue}) are split per the probe-C index (Appendix N.7); 10 of 100 cohort cases are skipped at probe-build time, so the C-row n is smaller than the others.

H Probe prompt templates

The shared system prompt and base user-prompt template are released in `contexts/system.txt` and `contexts/user_base.txt`. Probe-specific perturbations (A–G) are applied by `scripts/build_probe_prompts.py`: the Probe D boilerplate block, the Probe E drafting hints, and the Probe F/G retrieval-insertion templates are defined inline in that script.

I Five-channel metric computation

Reference implementations for channels C1 through C5 are provided in the supplementary material (`scripts/compute_c{1,2,3,4,5}.py`). Each channel is a deterministic function of generated text, parser output, and the XML gold post-amendment claim, as defined in §4.2.

J Attorney validation study

A 30–50-case double-rated sample assessing inter-rater agreement between our automated channels and licensed patent attorneys is committed post-review. Reported metrics will include Cohen's κ for categorical channels and Spearman ρ for continuous channels, against the automated scoring on the same cases. A 10-case pre-release sanity sample showed all four channels moving in the direction the reading attorney predicted; the full κ quantification is reserved for the published study.

The 10 cases used in the sanity sample are the same 10 cases across all four channels (a single shared subset, not a per-channel subset), drawn deterministically from the 100-case cohort by application-number ordering, so that each case contributes one C1, C2, C3, and C5 reading to the attorney's directional check. Application-number ordering correlates loosely with filing date and therefore with XML-format era, so the 10-case sanity sample is not stratum-balanced; the full 30–50-case study will instead draw a stratum-balanced sample across the six cohort axes so that the directional check is not era-biased. The reading attorney was blind to the automated scores at the time of directional prediction: per-case attorney readings were elicited from the input materials (pre-amendment claim, rejection rationale, cited prior art, and the model-generated amendment) before the automated scores were revealed; the comparison was then made between the attorney's predicted direction and the automated direction. The full κ study will pre-register the same blinding protocol on the larger 30–50-case sample, with double-rated readings to estimate inter-rater reliability alongside agreement with the automated channels.

K Batch replication protocol

Sampling procedure. Deterministic cohort selection (`scripts/select_cohort.py`) under the six-axis marginal-match procedure described in §5.1 and Appendix N. The procedure is reproducible from seed.

Batch 0. The seed used for primary analysis is 42; cohort case identifiers are released as `cohort_batch0.json`.

Batch 1 and 2 pre-specification. The same sampling criteria with different seeds are committed in `cohort_seeds.json` before experiments begin. Batches 1 and 2 are not drawn unless post-review scrutiny requests them.

Reviewer-request protocol. On request, batch 1 is drawn, LLM generations are run under identical conditions, and the primary H1, H2, and H3 ef-

fect sizes are recomputed. Consistency with batch 0 within the pre-specified equivalence margin is reported.

Why not pre-emptively. Primary results run against a single well-constructed cohort. Pre-emptive multi-batch analysis would either double the budget without informing the primary hypotheses or, if abbreviated, weaken probe depth. Batch replication is therefore reserved as a robustness mechanism, not a primary design.

K.1 Stage 2 escalation criteria

Stage 2 (Claude Haiku 4.5 and GPT-4o-mini on the full probe battery) is performed when at least two of the following hold consistently across both Stage-1 flagships: (i) Probe F yields $\Delta C5 > 0.2$ relative to no-retrieval baseline; (ii) Probes F and G yield comparable C5 shifts ($|\Delta_F - \Delta_G| < 0.1$, same sign); (iii) retrieval does not reduce C5 below baseline under at least one of F or G. When fewer than two conditions hold, Stage 1 already answers the question asymmetrically across vendor families, and Stage 2 primarily clarifies scope rather than shifting conclusions.

L Stratified results

Per-stratum result tables (by statute section, technology center, XML format, and amendment pattern class) are released as supplementary TSVs alongside the corpus and parsing code. The cohort's six-axis marginal match (Appendix N) ensures each stratum has at least five cases, supporting within-stratum effect-size estimation.

M Corpus release index

The released corpus consists of: the parsed JSONL of 7,385 four-tuples; the parsing code; the application-number index permitting zero-cost reconstruction from USPTO ODP; and `cohort_batch0.json` and `cohort_seeds.json` for batch replication. All released under a permissive open license; USPTO ODP data has no copyright restriction.

N Cohort construction

N.1 Why 100 cases

The cohort size is the smallest n that simultaneously satisfies four constraints: (i) within-model paired comparisons with >80% power at Cohen's $d = 0.3$, $\alpha = 0.05$; (ii) six-axis marginal-match feasibility with minimum cell size 5 (requires $n \geq \sim 90$); (iii) probe-depth budget at 7 probes $\times$ 3 reps $\times$ 4 models $= 84$ calls per case (8,400 calls before ablations) which is feasible at $n = 100$ but triples at $n = 300$ without adding decisive within-model power; (iv) attorney-validation headroom, since a 30–50-case audit is a meaningful fraction of $n = 100$ but only 3% of $n = 1{,}000$.

N.2 Why stratified, not random

A random 100-case draw from the 4,270-case eligibility pool would under-represent the rejection outcomes most informative for H2 contrasts (Reversed cells at ∼30% raw frequency), rare statutes (§101-related and §112-alone at ∼6–10% each, falling below the cell-minimum 5 threshold with non-trivial probability), and minority technology centers (long tail of 1–3-case bins that cannot support per-TC stratified robustness). Stratified sampling with target floors guarantees presence on all six axes.

N.3 Sampling procedure

Iterative marginal matching, not cross-product enumeration. The six-axis cross-product has too many empty cells at $n = 100$; iterative matching instead scores each case in the pool by the rarity of its six-axis vector relative to the targets, greedily accepts cases that move the cohort toward all six target marginals simultaneously, and uses the seed to break ties. At each step the procedure checks cell-minimum $m = 5$ for every bin of every axis. Code: `scripts/select_cohort.py`. Seed for batch 0: 42.

N.4 Eligibility filter

The 4,270-case pool is the subset of the 7,385 β-parsed corpus satisfying: (i) parse health (pre- and post-claims parseable, at least one CTNF rejection instance, statute section identifiable on the first rejection); (ii) axis coverage (subdecision outcome $\in$ {Affirmed, Reversed, Affirmed-in-Part}, excluding dismissed and remanded outcomes that appear in ∼8% of the source pool); (iii) no obvious data defect (pre-claim 1 $\geq$ 50 characters, at least one modified or cancelled per-claim action so that there is something to score).

N.5 Batch 0: achieved vs. target

Marginal match per axis: outcome 40/40 Affirmed, 40/40 Reversed, 20/20 AIP. Statute 35/20/15/15/10/5 target vs. 36/20/15/15/9/5 achieved across 103-alone, 102+103, 103+112, 101-related, 112-alone, and other. Technology

center 20/17/14/12/12 across the top six TCs (exact match). Pattern modify_only 45 / cancel_heavy 20 / mixed 15 / new_or_add 10 / other 10 (exact match). Format dtd_legacy 25 / ns_claims 75 (exact match). Year 25/45/30 target vs. 24/46/30 achieved across 2010–14, 2015–19, 2020–25. All bins pass the $m = 5$ minimum (smallest bin: statute "other" at 5).

N.6 Post-registration amendments

Cohort. PILOT-2019006614 was dropped because its β pre_clm XML contains two `<Claim>` wrappers with eleven unnumbered `<ClaimText>` elements and no inline-header pattern, making pre-amendment text not deterministically recoverable. A pre-declared replacement rule selected the lowest app_num case in the eligibility pool matching the dropped case's six-axis vector (Affirmed / §101-related / TC 3600 / 2020–25 / new_or_add / ns_claims) with parse-success under the patched parser. Selected: PILOT-2022000316. Marginal impact: zero.

Model matrix. GPT-4o-mini replaces GPT-5-mini in the smaller-OpenAI cell. Trigger: in the N=10 multi-model smoke run, GPT-5-mini produced 25 of 30 zero-parseable-claim responses, 13 of 30 max_tokens truncations, and did not accept the pre-registered temperature 0.3. A pre-declared fallback rule substitutes to the next stable smaller OpenAI model supporting temperature control and canonical claim format. Verification on GPT-4o-mini: 29 of 30 parseable responses, 0 of 30 truncations, temperature 0.3 accepted.

N.7 Retrieval-pool disjointness

The invariant cohort $\cap$ pool $= \emptyset$ holds by construction, since the pool is the β corpus with the 100 cohort cases removed.