GANDR: Grounded Answer Drafting with Claim-Level Verification, and Strict-Grounded Evaluation in the High-Stakes Legal Domain

Anonymous ACL submission

Abstract

In high-stakes domains such as legal practice, a language-model answer is only useful to the extent that a reader can verify each claim against the source the system cites. Current grounded-generation pipelines score the answer as a whole, asking whether it is faithful in aggregate rather than whether each individual claim is backed by the source attached to it. A correct conclusion can therefore rest on fabricated or loosely matched citations and still score well. Closing this gap requires both a system built for per-claim verification and an evaluation that measures it. We introduce GANDR (Grounded ANswer DRafter), a two-agent system in which a Drafter writes an answer in a structured legal-reasoning format. A separate Critic, with the same view as a human verifier, breaks the draft into individual claims and audits each against its cited source before the system commits. We pair the system with a strict correctness criterion under which every citation must resolve to a passage the retriever returned, extended across repeated runs to test stability. On a 185-item legal benchmark where all six systems share one backbone, one retrieval surface, and one citation instruction, GANDR ranks first on every primary metric, reaching 70.8 % strict accuracy and leading the strongest baseline by 11.3 points. Ablations attribute the lead to each architectural commitment, and the lead transfers to three further backbones spanning open- and closed-source ecosystems. The output of GANDR is a per-claim audit trace, not a single verdict. Code will be released upon acceptance.

1 Introduction

Legal AI is read by a verifier. When a lawyer uses an AI assistant to research a legal question, the work that follows the answer is verification, not exploration (Qian et al., 2026). The reader checks each claim against the source the system cites, asking whether the rule is actually stated in the cited case and whether the passage actually supports the argument. A single fabricated citation in a legal memo carries real professional cost. The operative question is not whether the model often reaches the right answer, but whether every claim is anchored in a source the reader can actually find. Audits put hard numbers on the failure mode. General-purpose LLMs hallucinate on 58 % to 88 % of open-domain legal queries (Dahl et al., 2024), and even retrieval-grounded commercial legal-AI products still do so on 17 % to 43 % (Magesh et al., 2025).

The verifier's question is the one current evaluation does not ask. The verifier asks whether each claim cites a real source and is supported by it, while evaluation today scores the answer in aggregate. Pipelines place source passages in context, instruct the model to cite them, and report a single faithfulness rate, citation-precision score, or correctness verdict (Liu et al., 2023a; Li et al., 2024; Gao et al., 2023). A correct conclusion can therefore rest on fabricated or loosely matched citations and still score well. However, a stricter metric alone is not enough. A system whose outputs can be verified claim by claim has to be built to produce that property, not patched with a post-hoc score. Standard remedies do not close the gap. Single-LLM generation drifts on citations even when the disposition is right (Niu et al., 2024; Liu et al., 2023a; Gao et al., 2023), and stacking more LLM calls in the usual multi-agent or self-correction style (Shinn et al., 2023; Madaan et al., 2023; Gou et al., 2024) can amplify rather than reduce the slip rate: each additional call is another chance to introduce a citation that does not resolve to retrieved authority.

Our approach. We propose GANDR (Grounded ANswer DRafter), a two-agent system that builds per-claim verification into generation (Figure 1). A Drafter produces an answer in a structured legal-reasoning format (Conclusion, Rule, Explanation, Application, Conclusion) (Qian et al., 2026) that

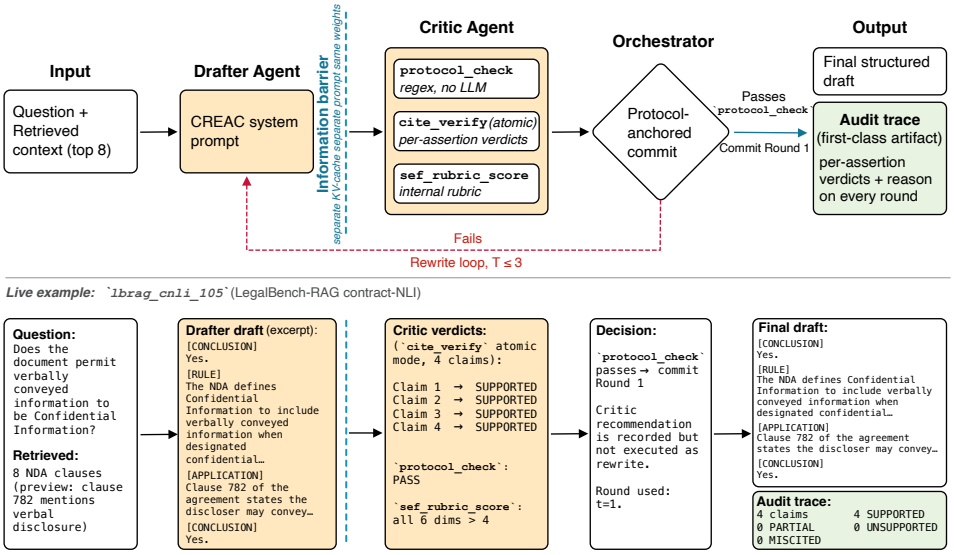

Figure 1: GANDR architecture (top) and a live trace (bottom). Drafter and Critic share backbone weights but run in *disjoint* LLM contexts: the Critic sees the question, retrieved passages, and the Drafter's final draft only, never the Drafter's chain of thought. The Critic invokes three tools (`protocol_check`, atomic `cite_verify`, `sef_rubric_score`) and writes per-assertion verdicts to the audit trace on every round. The orchestrator's commit rule is anchored to `protocol_check`. Bottom: live trace for `lbrag_cnli_105`.

lays every assertion out in the order a verifier would read it. A separate Critic then runs with the same view as a human verifier: it sees the question, the retrieved passages, and the final draft, but not the Drafter's internal reasoning. It breaks the draft into individual claims and audits each against its cited source before the system commits. The output is an answer paired with a per-claim audit trace. Alongside the system we score correctness *strictly*: an answer counts as correct only if it reaches the right disposition under an LLM-as-judge rubric and every citation it emits resolves to a passage the retriever actually returned. One fabricated or out-of-context citation fails the answer. We extend the criterion across repeated runs to test stability.

Evidence on a controlled cohort. We evaluate GANDR against five baselines ranging from zero-shot prompting to advanced RAG with reranking and multi-agent self-correction. All six systems share one self-hostable open-source backbone (Yang et al., 2026), one BM25 retrieval surface, and the same citation instruction, so observed differences reflect the system architecture rather than model scale, retrieval luck, or prompt compliance. On a 185-item benchmark drawn from LegalBench (Guha et al., 2023) and LegalBench-RAG (Pipitone and Alami, 2024), GANDR ranks first on every primary metric. It leads the strongest baseline by 11.3 points on strict accuracy, with the gap statistically significant at $p < 0.01$. Ablations attribute the lead to each architectural commitment, and the lead transfers across four backbones spanning open- and closed-source ecosystems.

Contributions. In summary, our work is threefold:

1. **A two-agent system that produces per-claim verification as a generation artifact.** A Drafter paired with a separate-context Critic emits an answer plus a per-claim audit trace on every round, and no baseline in standard configurations emits a comparable signal.
2. **A strict-grounded correctness criterion.** An answer counts as correct only if every citation it emits can actually be found in the retrieved sources, scored deterministically rather than judged in aggregate. We extend the criterion across repeated runs to stress-test stability.
3. **Evidence against five baselines and across four model backbones.** The lead is statistically significant against every baseline and transfers across open- and closed-source backbones, indicating it is a property of the architecture rather than of any one model.

2 Related Work

Faithfulness scoring for RAG. Faithfulness has been a central concern since the original RAG architecture (Lewis et al., 2020), with hallucination documented at the system level (Huang et al., 2025; Gao et al., 2024). Rashkin et al. (2023) formalized the Attributable-to-Identified-Sources framework, and a growing line then documented that LLMs routinely produce answers whose disposition is

correct but whose citations are unreliable, post-rationalized, or unsupported (Menick et al., 2022; Bohnet et al., 2023; Gao et al., 2023; Niu et al., 2024; Liu et al., 2023a; Li et al., 2024; Wallat et al., 2025). Wallat et al. (2025) name post-rationalized citation as the failure mode and Gao et al. (2023) introduce the ALCE benchmark for citation-aware generation. The metrics in this line almost always aggregate at the answer level (a faithfulness rate, a citation-correctness score, a single binary verdict). GANDR makes per-claim verification both the production target and the evaluation target, decomposing every draft into atomic $(\text{claim}, \text{rule_id}, \text{span})$ tuples and emitting a per-assertion verdict on every round.

Beyond single-score leaderboards. Single-score leaderboards are sensitive to choices the evaluator leaves implicit. Peyrard et al. (2021) re-evaluate hundreds of NLP setups and find aggregation choice alters the top system in roughly thirty percent of them. Opitz (2024) document that metric choices in shared tasks are frequently unjustified. Our evaluation is a constructive response: a strict-grounded conjunction beside the lenient column, extended across repeated runs as Stable & Correct @ k.

Structured prompting and agent frameworks. SEF (Qian et al., 2026) showed that a CREAC schema with a self-rubric produces analyses more legible to practitioners than free-form prompting. We treat SEF-prompted single-LLM generation as the head-to-head reference (B2, runner-up on strict). A broader family of self-correction methods runs the reviewer in the *same* context as the generator: ReAct (Yao et al., 2023), Reflexion (Shinn et al., 2023), Self-Refine (Madaan et al., 2023), Chain-of-Verification (Dhuliawala et al., 2024), and CRITIC (Gou et al., 2024). All rely on same-context self-criticism, the architectural choice GANDR rejects. CrewAI (CrewAI, 2026) (B4) and a LangGraph (Langchain, 2026) self-correction graph (B5) cover generic role-based delegation and same-context self-correction. B5 vs. GANDR is the like-for-like contrast.

LLM-as-judge. Pairwise and rubric-based LLM-as-judge protocols (Zheng et al., 2023; Liu et al., 2023b) are standard, and a parallel line documents biases including position effects and self-preference for in-family generations (Wang et al., 2024; Panickssery et al., 2024). Two failure modes drive our setup. We address self-preference bias by a **cross-lineage split** (NVIDIA generator vs. Anthropic judge), and we address unverified judge fidelity with calibration against two independent annotators (§5.7).

Legal NLP benchmarks and scope. LegalBench (Guha et al., 2023) and LegalBench-RAG (Pipitone and Alami, 2024) provide the public, domain-stratified surfaces from which our benchmark is drawn. Closed commercial legal-AI products (Lexis+ AI, Harvey, CoCounsel) are not reproducible, vary across releases, and do not expose the per-claim audit shape we measure, so a direct comparison would conflate model, system, and product effects and is out of scope.

3 The GANDR System

GANDR is a two-agent system that builds claim-level verification into generation. Three architectural commitments distinguish it from a conventional grounded-generation pipeline: (i) a CREAC-shaped generation schema lays every claim out for inspection; (ii) a Critic agent runs in a separate LLM context and audits each claim via atomic-claim decomposition; and (iii) an orchestrator anchors the commit rule to a structural protocol check, not to the Critic's PASS/REWRITE verdict. Each commitment is a subtractive target in §5.

3.1 CREAC as a verification contract

CREAC (Conclusion, Rule, Explanation, Application, Conclusion) is a *reading* order, not a writing order. Practitioners scan in CREAC order to verify, and the writer's job is to produce text that survives that scan. We treat CREAC as a contract with three formal commitments. **Block presence and order:** five blocks `[CONCLUSION]` `[RULE]` `[EXPLANATION]` `[APPLICATION]` `[CONCLUSION]` in that order, both `[CONCLUSION]` blocks non-empty. **Inline cites grounded in the retrieved corpus:** each `{cite: <rule_id>}` must reference a passage in the retrieved context. **Per-assertion accountability:** every claim in `[EXPLANATION]` or `[APPLICATION]` must pair with a specific cited rule and be independently auditable. The first commitment is enforced by a regex-based `protocol_check` that contains no LLM (cheap, objective, ungameable). The second and third are enforced by atomic-claim `cite_verify`.

3.2 Drafter and Critic in disjoint contexts

The Drafter is an LLM call against the shared backbone with a CREAC-shaped system prompt and the retrieved top-8 passages (§4). It emits a CREAC-

structured draft and is *not* told the Critic's pass criterion (its job is structural well-formedness, not anticipating the critique).

The Critic runs in a *separate* LLM context against the same backbone weights. "Separate context" means a fresh KV-cache, a distinct system prompt, and a restricted input. The input consists only of the Drafter's final draft text, the question, and the retrieved passages, never the Drafter's chain of thought. The architectural commitment is the *information barrier* between Drafter reasoning and Critic input. The Critic's view matches a human verifier's. The Critic invokes three tools: (1) **`protocol_check`**, regex-only, returns `{passed, missing_blocks, order_violation}`; (2) **`cite_verify`** (atomic, default), which splits `[EXPLANATION]/[APPLICATION]` into (claim, rule_id, span) tuples and verdicts each as `SUPPORTED/PARTIAL/UNSUPPORTED/MISCITED` (a holistic variant exists only for the ablation). The third tool, **`sef_rubric_score`**, is a six-dimension 1–5 draft-quality rubric inherited from SEF (Qian et al., 2026) covering answer placement, clarity, conclusion isolation, terminology consistency, claim–evidence alignment, and fact specificity, scored by the local generator at $T{=}0$ (full rubric in Appendix D.6). Only the first two tools enter the headline. SEF is internal Critic machinery used to phrase actionable rewrite directives when the protocol gate fails.

Why two contexts, not one. A single-context loop (Reflexion-style, B5) gives the model both the reasoning that produced the draft and the obligation to critique it. Two failure modes follow: *anchoring*, where the second pass ratifies the first pass's framing because the KV-cache contains both, and *rubric-stamping*, where the model emits rubric-positive language without earning it because prompt and rubric share a context. Separating contexts breaks both, and §5 quantifies the gap.

Why atomic, not holistic. A holistic "does this draft cite faithfully?" check rubber-stamps drafts that contain a few accurate cites and many unsupported claims, leaving the audit log without actionable signal. Atomic mode forces per-assertion verdicts, turning the trace into a list of claim-cite-verdict triples a human can scan in seconds.

3.3 Protocol-anchored orchestrator

The Critic emits per-assertion verdicts and rewrite directives, but the orchestrator's commit rule is anchored to the structural `protocol_check`, not to the Critic's PASS/REWRITE verdict. **Round-1 anchor (default path):** if `protocol_check` passes on the first draft, the orchestrator commits Round 1, and the Critic still runs and writes verdicts to the audit trace but acts as an *auditor*. **Rewrite loop (fallback):** if `protocol_check` fails on Round 1, the Critic's directives are prepended to the Drafter's next-round input, up to $T{=}3$ rounds. On exhaustion the system returns `passed=False` honestly. There is no fall-through to a "best effort" PASS. The atomic-verification threshold inside the rewrite path (0.50) is corpus-calibrated. The "−Round-1 anchor" ablation in §5 reverts the threshold to the original 0.95 value to isolate whether the threshold or the protocol-anchored commit itself is the load-bearing element.

4 Experimental Setup

Benchmark. A 185-item benchmark drawn from LegalBench (15 task tags: `hearsay`, `learned_hands` (3 sub-tasks: `courts`, `crime`, `employment`), `nys_judicial_ethics`, `definition_extraction`, `overruling`, `scalr`, `ssla`, `rule_qa`, `citation_prediction`, `consumer_contracts_qa`, `corporate_lobbying`, `personal_jurisdiction`, `maud`) and LegalBench-RAG (contract-NLI, CUAD). Per-bucket quotas: Appendix D.1.

Retrieval surface. A BM25 index of 13,090 passages from inline contexts and the LegalBench-RAG mirror. Every system receives the same top-8, so stability measures *system* behavior. We frame BM25 as a *standardized context-delivery mechanism*, not a retrieval contribution. For the bulk of LegalBench items, BM25's top-8 is dominated by passages from the item's own inline context. Genuine external retrieval is exercised only on the LegalBench-RAG cross-domain slice ($n{=}30$). The architectural lever we measure is production-time per-claim audit, not retrieval quality. B3 (advanced RAG) varies retrieval rather than generation, so the B3-vs-GANDR contrast tests whether the lever could be a retrieval-quality artifact.

Systems. Six systems share one backbone and one retrieval surface. **B1** zero-shot rules out "the model already knows law"; **B2** is a CREAC-prompted single LLM (SEF-style), runner-up on strict; **B3** is advanced RAG (BM25 + dense + cross-encoder rerank); **B4** is a CrewAI (CrewAI, 2026) Researcher+Writer crew; **B5** is a LangGraph (Langchain, 2026) self-correction graph

System	Lenient	Strict	Δ (strict−lenient)	S&C @ 5	CREAC
B1 zero-shot	57.3	38.4	−18.9	0.049	0 %
B2 SEF prompt	60.0	59.5	−0.5	0.297	100 %
B3 advanced RAG	60.5	43.2	−17.3	0.049	0 %
B4 CrewAI	58.4	21.6	−36.8	0.000	0 %
B5 LangGraph	51.4	21.6	−29.8	0.000	0 %
B6 GANDR	**71.9**	**70.8**	**−1.1**	**0.460**	**100 %**

Table 1: Headline metrics on the 185-item benchmark, accuracy in percent. Every system carries the same shared GROUNDING REQUIREMENT instruction (§4). *Lenient*: judge ≥ 4 on the operative holding only. *Strict*: *Lenient* and the cite-validity check. Δ: strict−lenient, small for systems that internalize grounding (GANDR, B2), large for the others. 95 % Wilson CIs on strict: B1 $[31.7, 45.6]$, B2 $[52.3, 66.3]$, B3 $[36.3, 50.4]$, B4 $[16.3, 28.1]$, B5 $[16.3, 28.1]$, GANDR $[63.9, 76.9]$. On S&C, GANDR $[0.389, 0.531]$ vs. B2 $[0.236, 0.367]$, and GANDR's S&C CI does not overlap any baseline. On strict, GANDR's CI is disjoint from B1/B3/B4/B5, and the small overlap with B2's upper bound (66.3 % vs. GANDR's lower bound 63.9 %) is consistent with paired McNemar at $p < 0.01$.

(same context); **B6** is GANDR. B2 vs. B6 is the head-to-head, and B5 vs. B6 is the single-context vs. separate-context contrast.

Citation-instruction parity. Every system carries the same shared GROUNDING REQUIREMENT block in its primary system prompt, instructing the model (i) to cite only identifiers that appear verbatim as `rule_id` headers in the retrieved context, and (ii) to follow every claim with an inline `{cite: rule_id}` marker. This parity makes strict scoring a *content* check (did the model anchor in retrieved authority?) rather than a *compliance* check. Per-system overlays: Appendix D.5.

Backbone and judge. All six systems run on NVIDIA **Nemotron-Cascade-2-30B-A3B** (FP8, vLLM on a single H100 NVL), a Mamba2-Transformer hybrid MoE with 30B total / 3B active per token. A single self-hostable backbone rules out the "GANDR wins because it uses a bigger model" rebuttal. The judge is **Claude Opus 4.7** (Anthropic, 2026), invoked as a Claude Code sub-agent. The cross-lineage split (NVIDIA generator vs. Anthropic judge) blocks self-preference bias structurally. To check the lever is not Nemotron-specific, we re-run the B2-vs-GANDR head-to-head on three further backbones spanning two ecosystems: GLM-4.7-Flash (open-source), `gpt-5.4-mini`, and `gpt-5.4` (§5.5).

Metrics. Three primary columns plus one diagnostic. (**A**) **Lenient Accuracy**: the conventional disposition-only judge bit (rubric $\geq$ 4). (**B**) **Strict Accuracy**: lenient *and* a deterministic grounding check requiring (i) at least one inline `{cite: rule_id}` marker and (ii) every extracted `rule_id` resolves to a header in the retrieved top-8. A single fabricated or out-of-context citation fails the check. The strict-minus-lenient gap is the diagnostic of whether grounding is internalized as architecture or merely correlated with through prompting. (**C**) **Stable & Correct** @ k=5: judge-equivalent across all $C(5, 2)$=10 pairs, strict-correct on the representative answer, and strict-grounded on every one of the 5 runs. Stability is judge-anchored rather than string-based so equivalent paraphrases are not penalized. (**D**) **CREAC compliance** (fraction passing `protocol_check`): transparency only.

Protocol. k=5, T=3 maximum Critic rounds, atomic-verification threshold 0.50 (ablation at the original 0.95). The rewrite loop is a robustness safety mechanism. In the headline run, 98.6 % of items commit on Round 1 (Appendix C). All systems use temperature 0.7 and `max_tokens` = 2048.

5 Results

5.1 Headline

Table 1 reports the headline. GANDR leads on every primary column: lenient (71.9 %), strict (70.8 %), and S&C @ 5 (0.460). The gap over B2 is 11.3 pp on strict and 0.163 on S&C, both clearing McNemar at $p < 0.01$. Every other baseline trails at $p < 10^{-8}$ on strict (Appendix A.1, Table 6).

Without a structured conclusion, free-form generation drifts. B1 (zero-shot) and B3 (advanced RAG) emit free-form prose. Both carry the same GROUNDING REQUIREMENT, but neither pins the conclusion to a labeled block or ties each claim to a cite slot. Lenient is competitive (57.3 %, 60.5 %) while strict collapses to 38.4 % and 43.2 %, because the model emits citations in a non-canonical bracket family the verifier rejects, or omits the citation entirely (Appendix B.1 for verbatim examples).

CREAC structure absorbs the drift. B2 fixes backbone, retrieval, and citation instruction and

Item: lbrag_cuad_2886 **Round: 1** **protocol_check: PASS** **Drafter draft committed**

Question: "Does the Agreement address ownership of intellectual property arising from the co-marketing relationship?"

Atomic claim	rule_id	Verdict	Reason
The Agreement governs the parties' co-marketing arrangement.	`corpus: lbrag_cuad_886_clause_1`	SUPPORTED	*Clause 1 explicitly defines the co-marketing scope.*
Termination requires 60 days' written notice.	`corpus: lbrag_cuad_886_clause_7`	SUPPORTED	*Clause 7 specifies the 60-day notice term verbatim.*
Either party may terminate for material breach.	`corpus: lbrag_cuad_886_clause_7`	PARTIAL	*Clause 7 mentions breach but does not specify "either party".*
The Agreement permits assignment with consent.	`corpus: lbrag_cuad_886_clause_12`	SUPPORTED	*Clause 12 directly addresses assignment with prior written consent.*
The Agreement does not address ownership ofintellectual property.	`corpus: lbrag_cuad_886_clause_9`	UNSUPPORTED	*Clause 9 covers revenue sharing, not IP ownership; topic mismatch.*
Disputes are governed by California law.	`corpus: lbrag_cuad_886_clause_15`	SUPPORTED	*Clause 15 specifies California as the choice of law.*

Legend: SUPPORTED The cited rule supports the claim. PARTIAL The rule partially supports the claim. UNSUPPORTED The cited rule does not support the claim. MISCITED The cited rule is the wrong rule.

Figure 2: Atomic-claim audit trace from GANDR on `lbrag_cuad_2886`: six per-claim verdicts. The `UNSUPPORTED` verdict on claim 5 catches a topic mismatch, and `protocol_check` still passes so the trace surfaces the under-supported claim rather than blocking commit.

adds the CREAC five-block schema. Every claim in `[EXPLANATION]`/`[APPLICATION]` must end with an inline `{cite: rule_id}`, and the strict-lenient gap collapses to 0.5 pp. B2 lifts strict by 21.1 pp over B1 and 16.3 pp over B3 on the same backbone, retrieval, and instruction.

More agent calls without auditing amplify hallucination. B4 (CrewAI) and B5 (LangGraph) add multiple LLM calls without per-claim auditing. Lenient does not improve over B2 (58.4 %, 51.4 %), and strict drops to 21.6 % on both. On B4 the Researcher feeds the Writer fabricated rule-label strings. On B5 the single-context Plan-Draft-Review-Rewrite loop preserves `rule_id` corruptions across rounds (Appendix B.1). Each extra call is another chance to introduce a citation that does not resolve, and cite-validity fails on one slip. Across 925 runs, B4 and B5 produce no item whose all-five-seeds are simultaneously strict-grounded, so S&C@5 collapses to 0 (§5.3).

GANDR layers an audited Critic on top of CREAC. GANDR inherits B2's CREAC Drafter verbatim and adds three commitments: a Critic in a separate LLM context, atomic per-claim `cite_verify`, and a protocol-anchored commit. The Critic catches the citation slips a CREAC-prompted single LLM still produces under sampling noise, and the protocol-anchored commit prevents the multi-call amplification that B4 and B5 pay. Strict lifts 11.3 pp over B2 (70.8 % vs. 59.5 %) and S&C@5 reaches 0.460, 0.16 above B2 and at least 0.41 above every non-B2 baseline.

Variant	Lenient	Strict	Δ Strict
GANDR (full)	**71.9**	**70.8**	—
− Round-1 anchor	49.2	48.1	−22.7
− Atomic `cite_verify`	54.1	54.1	−16.7
− Critic	58.9	58.4	−12.4

Table 2: Subtractive ablations on GANDR's three commitments, accuracy and drops in percent. − *Atomic* `cite_verify`: replaces atomic with a holistic support check. − *Round-1 anchor*: reverts to the original strict-gate orchestrator. − *Critic*: Drafter alone under the protocol-anchored commit. Every variant keeps the near-zero strict-lenient gap ($\leq$ 1.1 pp), so the levers act on the *level* of accuracy, not the *shape* of the strict-lenient relationship. k=5, same five seeds as full GANDR.

5.2 Ablations

Table 2 isolates the three commitments. **Round-1 anchor** (−22.7 pp): the original strict gate (atomic $\geq$ 0.95 AND rubric mean $\geq$ 4.0) exceeds the verifier's reliable signal on this corpus ($\sim$ 0.30), so rewrites fire on $\geq$ 95 % of items and quality degrades across rounds. **Atomic `cite_verify`** (−16.7 pp): the holistic fallback emits a single "mostly supported" verdict that the audit log cannot act on. **Critic** (−12.4 pp): running the Drafter alone under the protocol-anchored commit still costs 12.4 pp. The Critic's contribution is real but smaller than the other two, consistent with the design (the protocol anchor is the load-bearing choice, and the Critic's distinct role is to produce the audit artifact).

5.3 Auditability

Figure 2 shows what GANDR ships alongside each answer: an atomic-claim audit trace with one ver-

Ecosystem	Backbone	System	Strict	S&C	CREAC	Δ vs. B2
open-source	Nemotron-Cascade-2	B2 SEF	59.5	0.297	100.0	—
	Nemotron-Cascade-2	**GANDR**	**70.8**	**0.460**	**100.0**	**+11.3**
	GLM-4.7-Flash	B2 SEF	62.7	0.411	96.8	—
	GLM-4.7-Flash	**GANDR**	**69.2**	0.297	**100.0**	**+6.5**
closed-source	`gpt-5.4-mini`	B2 SEF	74.1	0.638	100.0	—
	`gpt-5.4-mini`	**GANDR**	**77.8**	**0.686**	**100.0**	**+3.7**
	`gpt-5.4`	B2 SEF	76.8	0.692	100.0	—
	`gpt-5.4`	**GANDR**	**80.0**	**0.762**	**100.0**	**+3.2**

Table 3: Backbone-pluggability check. Strict accuracy and CREAC compliance in percent, strict S&C @ 5 as a proportion in $[0, 1]$, for GANDR and B2 SEF across four backbone panes. The S&C column matches the format in Table 1. All four panes use the same judge protocol, the same cite-validity strict criterion, and the same shared GROUNDING REQUIREMENT citation-instruction block. The Nemotron values match Table 1. The Δ column is the point estimate of GANDR −B2 in strict accuracy. Strict accuracy and S&C@5 track different properties (per-answer cite-validity vs. all-five-seeds cite-validity). GLM is the one pane where GANDR leads on strict but trails on S&C@5.

dict per claim. On this item, claim 5 is flagged `UNSUPPORTED` because its cited passage, though in the retrieved top-8, is only topically adjacent. Strict accuracy would not catch this. It checks that every citation resolves to a retrieved passage, not whether the passage actually supports the claim. The audit is the orthogonal signal, and the only such signal any system in the cohort emits. It matters most under larger or noisier retrieval pools, where topically-adjacent passages are easier to land on.

Shape of the artifact GANDR produces. Every GANDR answer ships with a (claim, rule_id, verdict, reason) list aligned to `[EXPLANATION]`/`[APPLICATION]`. The verdict mix at seed 0 (191 rounds, 737 atomic assertions): `SUPPORTED` 34.7 %, `PARTIAL` 30.5 %, `UNSUPPORTED` 18.5 %, `MISCITED` 16.3 %. The trace is emitted on every round (mean t_{used}=1.03).

What the audit catches that strict does not. At seed 0, GANDR passes strict on 131/185 items. The audit flags at least one `MISCITED`/`UNSUPPORTED` claim on 63 (48.1 %) of these and at least half the claims on 42 (32.1 %). Strict checks cite resolution. The audit verifies whether each cited passage supports its claim, which matters when the retrieval pool is large enough that topically-adjacent passages can be cited as if directly on point.

Why B4 and B5 collapse to 0 on S&C @ 5. Despite non-trivial strict (21.6 % each), B4 and B5 collapse to 0 on S&C@5. Across 925 B4 runs, 27.2 % emit *no* `{cite:}` marker and 35.5 % emit at least one unresolved cite, leaving only 37.3 % that pass the cite-validity check on that single run (B5: 14.7 %/40.6 %/44.6 %). S&C@5 requires cite-validity to pass on *every* one of 5 seeds for the same item, so only 8 B4 and 9 B5 items clear that prefix, and intersections with the judge-equivalence subset are empty.

5.4 Per-domain breakdown

GANDR leads on 7 of 8 domains in S&C@5 and 6 of 8 on strict (Appendix Figure 3), with B2 as the only competitive baseline. Largest leads over B2: employment (+0.55), criminal procedure (+0.25), contract (+0.16), federal procedure (+0.12), tort (+0.12). Two strict losses: securities (0.85 vs. B2 0.90, one item) and constitutional (0.55 vs. B2 0.80). The single S&C loss is constitutional (0.30 vs. 0.35), driven by classification items where the Drafter produces a substantive CREAC memo when a category label was the target, a schema-supervision mismatch we flag as future work.

5.5 Backbone pluggability

To confirm the lever is not Nemotron-specific, we re-run the B2-vs-GANDR head-to-head on three further backbones spanning two ecosystems: a second self-hostable open-source model (GLM-4.7-Flash, a 30B / 3B-active MoE in a different model family from Nemotron-Cascade-2) and two closed-source commercial models from a different vendor (OpenAI's `gpt-5.4-mini` and `gpt-5.4`). All four panes hold every other component constant: the same Drafter system prompt, the same Critic tools, the same BM25 top-8 retrieval surface, the same 185-item benchmark, the same Claude Opus 4.7 judge in the same chunked configuration, and the same shared GROUNDING REQUIREMENT block carried by every system in the cohort. Only the generator backbone changes.

GANDR retains the strict-accuracy lead over B2 on every backbone (Table 3): +11.3 pp on Nemotron, +6.5 pp on GLM-4.7-Flash, +3.7 pp

on `gpt-5.4-mini`, and $+3.2$ pp on `gpt-5.4`, with CREAC compliance at $100\,\%$ on every pane. GANDR on `gpt-5.4-mini` ($77.8\,\%$ strict) edges past B2 on the flagship `gpt-5.4` ($76.8\,\%$), so wrapping the cheaper backbone in GANDR's commitments yields a larger lift than upgrading the backbone under B2. The commitments are properties of the orchestrator, while absolute strict level still tracks backbone capability.

Audit shape is portable across backbones. Beyond accuracy, the orchestrator produces the same architectural signature under every pane. The protocol-check pass-rate is effectively $100\,\%$ on Round-1 drafts on all four backbones, and atomic-claim decomposition produces a consistent shape across families (mean 4.4 claims per draft on Nemotron, 4.3 on GLM, 5.4 on `gpt-5.4-mini`, 8.6 on `gpt-5.4`). The architectural commitments are properties of the orchestrator and the Critic tools, not of any particular model's weights or ecosystem.

5.6 Cost–accuracy tradeoff

Table 4 lists cost-accuracy operating points on the two closed-source backbones, holding cohort, benchmark, retrieval, and judge constant (per-call breakdown in Appendix B.2).

Backbone	System	Strict	Cost
`gpt-5.4-mini`	B2 SEF	74.1 %	∼$0.85
`gpt-5.4-mini`	**GANDR**	**77.8 %**	∼$2.45
`gpt-5.4`	B2 SEF	76.8 %	∼$3.10
`gpt-5.4`	**GANDR**	**80.0 %**	∼$10.30

Table 4: Strict accuracy and order-of-magnitude API spend per single-seed run on the two closed-source backbones (OpenAI list prices 2026-05-22, with empirical per-call token usage from k=5 artifacts, Appendix B.2). Diagonal: GANDR on cheaper backbone edges past B2 on flagship at lower spend.

The diagonal comparison. GANDR on `gpt-5.4-mini` reaches $77.8\,\%$ strict at ∼$2.45, edging past B2 SEF on the flagship `gpt-5.4` ($76.8\,\%$ at ∼$3.10) on both axes. Mini-with-audit lands within 2.2 pp of GANDR on the flagship ($80.0\,\%$ at ∼$10.30) for $\sim 24\,\%$ of the flagship cost. The audit costs $\sim 3\times$ in tokens over the single-LLM baseline on the same backbone, but shifts what a fixed budget buys: a mini-model answer with an atomic audit trace, rather than a flagship prompt-only answer.

5.7 Judge calibration

To rule out judge idiosyncrasy, we calibrate Claude Opus 4.7 against two independent annotators applying the literal rubric on n=120 outputs (20/system, stratified by judge correctness band) via a blinded HTML interface (system identity and judge score stripped). Pre-registered: Pearson $r \geq 0.80$, weighted Cohen's $\kappa_q \geq 0.60$.

Metric	A1 ↔ J	A2 ↔ J	Pooled
Pearson r	0.856	0.837	0.846
Weighted κ_q	0.854	0.835	0.845
Exact agreement	78.3 %	77.5 %	77.9 %
Within-1 agreement	89.2 %	88.3 %	88.8 %
Inter-annotator r, κ_q	0.979, 0.979		

Table 5: LLM-judge calibration against two independent annotators (n=120). Both annotators clear pre-registered $r \geq 0.80$ and $\kappa_q \geq 0.60$. Inter-annotator κ_q=0.979. Per-system and per-judge-bin breakdowns: Appendix A.2.

Both annotators clear the thresholds (Table 5; pooled r=0.846, κ_q=0.845). Per-system κ_q ranges 0.74–0.99, all clearing 0.60. The single structural gap is the 4-band: on 9 items the judge rates 4, both annotators round up to 5. Collapsing to $\{1, 2, 3, 4{=}4|5\}$ lifts exact agreement from $78\,\%$ to $86\,\%$ without changing any pass/fail.

6 Conclusion

Grounded generation in the high-stakes legal domain needs answers that can be verified claim by claim, not just answers that score well in aggregate. GANDR is a two-agent system that builds this verification into generation: a Drafter writes a CREAC-structured answer, and a separate Critic breaks the draft into individual claims and audits each against its cited source before the orchestrator commits. We scored correctness strictly under a cite-validity criterion and extended it across repeated runs to test stability. On a 185-item benchmark, GANDR ranked first on every primary metric and led the strongest baseline on strict accuracy by 11.3 points ($p < 0.01$). Two regularities anchor this. Free-form generation drifts on citations even when the disposition is right, and the CREAC schema absorbs the drift. Stacking more LLM calls amplifies the drift unless audited, and GANDR's separate-context Critic provides the audit without paying the multi-call hallucination tax. The lead survived three ablations (12 to 23 points each) and a four-backbone pluggability check across open- and closed-source ecosystems. The output is a per-claim audit trace, not a single verdict.

7 Limitations

Scope and setting. The system and evaluation are documented on one 185-item English-language U.S. legal benchmark, with CREAC as the sole headline schema, and the generalization argument is structural rather than a multi-domain replication. External retrieval is exercised only on the LegalBench-RAG slice (n=30). For the bulk of items BM25 returns passages from the item's own attached context, so retrieval is held constant as a context-delivery surface.

Backbone and judge. Headline on Nemotron-Cascade-2-30B-A3B. Three pluggability backbones add $+11.3/+6.5/+3.7/+3.2$ pp (§5.5), with S&C@5 trailing B2 on GLM even though strict leads. The Claude Opus 4.7 judge with cross-lineage split limits but does not eliminate single-source bias, and calibration shows r=0.85, κ_q=0.85.

Threshold and cost. The 0.50 atomic-verification threshold is corpus-calibrated, and the protocol-anchored commit (independent of the threshold) is the load-bearing choice. GANDR spends $\sim$ 5$\times$ more tokens than B1/B2 (Appendix B.2).

Risks. GANDR detects citation hallucinations that would otherwise reach the reader (§1, §5.3). The work introduces no new sensitive data and releases no new model weights.

References

Anthropic. 2026. Introducing claude opus 4.7. https://www.anthropic.com/news/claude-opus-4-7. Retrieved: May 22, 2026.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, and 3 others. 2023. Attributed question answering: Evaluation and modeling for attributed large language models. *Preprint*, arXiv:2212.08037.

CrewAI. 2026. The leading multi-agent platform. https://crewai.com. Retrieved: May 22, 2026.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, page 3563–3578. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. Critic: Large language models can self-correct with tool-interactive critiquing. *Preprint*, arXiv:2305.11738.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Langchain. 2026. Langgraph: Agent orchestration framework for reliable ai agents. https://www.langchain.com/langgraph. Retrieved: May 22, 2026.

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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2024. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 493–516, Bangkok, Thailand. Association for Computational Linguistics.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *Preprint*, arXiv:2203.11147.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2026. Introducing gpt-5.4 mini and nano. `https://openai.com/index/introducing-gpt-5-4-mini-and-nano/`. Retrieved: May 22, 2026.

Juri Opitz. 2024. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 12:820–836.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of NLP systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2301–2315, Online. Association for Computational Linguistics.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *Preprint*, arXiv:2408.10343.

Chen Qian, Yimeng Wang, Yu Chen, Lingfei Wu, and Andreas Stathopoulos. 2026. From "thinking" to "justifying": Aligning high-stakes explainability with professional communication standards. *Preprint*, arXiv:2601.07233.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 22–32, New York, NY, USA. Association for Computing Machinery.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, Renjie Pi,

Grace Lam, Nayeon Lee, Alexander Bukharin, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2026. Nemotron-cascade 2: Post-training llms with cascade rl and multi-domain on-policy distillation. *Preprint*, arXiv:2603.19220.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Judge reliability and headline significance

A.1 Per-domain breakdown and McNemar significance triples

Full per-domain strict-accuracy and Stable & Correct numbers and the appendix-only diagnostic metrics (Stable & Wrong @ k, citation-set Jaccard, exact-string conclusion match) are released with the artifact bundle.

A.2 Judge calibration: per-system and per-bin breakdowns

The main-body judge-calibration result in §5.7 is anchored by the per-system and per-judge-bin breakdowns reported here. The 120-item set is 20 items per system across 6 systems, with the per-system sub-sample stratified by judge correctness band to match that system's full-benchmark distribution. Two independent human annotators apply the literal judge rubric through a blinded HTML interface, with no access to system identity, the LLM-judge's score, or each other's ratings. Both annotators are research collaborators on this project who consented to the use of their ratings.

Per-system breakdown. Per-system κ_q, Pearson r, and means against the judge are in Table 7. Every system clears the per-system $\kappa_q \geq 0.60$ pre-registered threshold. The range is 0.74 (B6 GANDR vs. A2) to 0.99 (B3 vs. both annotators). The largest mean delta is B5, where annotators land $+0.75$ above the judge: agreement on absolute level differs on a few B5 items while ranking remains consistent.

Per-judge-bin breakdown. Calibration broken down by the judge's own correctness band is in Table 8. On the 50 items the judge rates 1, the annotators agree exactly 84–86 % of the time and within one band 92 %. On the 49 items the judge rates 5, agreement is 92 % exact and 100 % within one. The single structural gap is the 4-band. On the 9 items the judge rates 4, both annotators round up to 5. This is a band-anchor disagreement on a specific rubric transition, not a systematic judge bias.

A.3 Annotation interface for the calibration set

The 120-item judge-calibration set described in §5.7 and Appendix A.2 was rated through a custom blinded HTML interface. The interface strips system identity and the LLM-judge's score from the client-side payload. Figure 4 shows the rating screen the two annotators saw.

B Cohort and backbone diagnostics

B.1 Why B1, B3, B4, B5 drop strict accuracy

The cite-validity strict criterion has two parts: at least one inline `{cite: rule_id}` marker is present, and every emitted marker resolves to a `rule_id` that appears verbatim in the item's retrieved top-8 header set. An answer that meets both passes, and an answer that fails either is strict-wrong. The four single-LLM and agent baselines fail this check in qualitatively different ways, summarized in Table 9 and elaborated below with verbatim seed-0 outputs.

B1 zero-shot (-18.9 pp). B1 fails almost exclusively because the model emits no canonical `{cite: rule_id}` marker on a sub-population of items, despite the GROUNDING REQUIREMENT block instructing it to do so. The model still references the source material, but in alternative bracket families the regex rejects. Three representative surface forms:

- **Non-ASCII brackets** (`lbrag_cnli_105`): the model wraps the `rule_id` in U+3010 / U+3011 fullwidth brackets, a Unicode encoding error from the model's perspective. The marker is not rendered here.
- **Square brackets** (`lbrag_cuad_2886`): "The only passage that addresses IP ownership is in `[corpus:lbrag_cuad_15]`, which states..."
- **Braces without `cite:` prefix** (`lbrag_cuad_3056`): "...a new agreement is signed that invalidates it. `{lbrag_cuad_2328}`"

B1's disposition reasoning is sound and the `rule_id` is the right one. The bracket family is

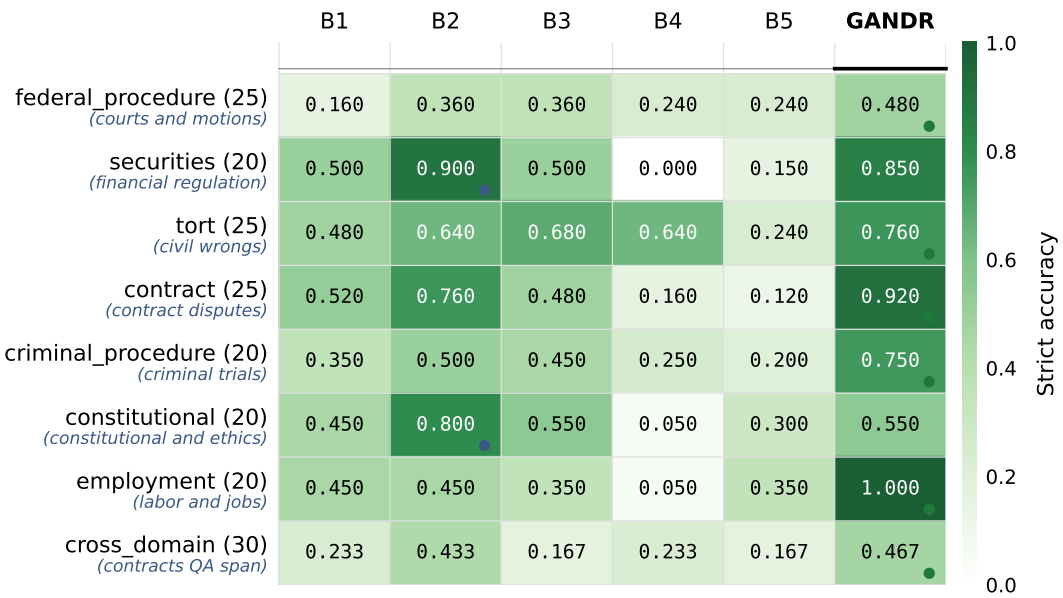

Figure 3: Per-domain *strict accuracy* (cite-validity criterion) on the eight retained buckets. Cells show the strict-correct rate per system at the representative seed, and the green dot in each row marks the per-row leader. GANDR leads on 6 of 8 domains, with the largest gaps clustering on long-form doctrinal-application items (employment $+0.55$, criminal procedure $+0.25$, contract $+0.16$, federal procedure $+0.12$, tort $+0.12$ vs. B2). The two strict losses are securities (n=20, 0.85 vs. B2 0.90, a one-item gap) and constitutional (n=20, 0.55 vs. B2 0.80). The longitudinal Stable & Correct @ 5 view (not plotted here) follows the same pattern with the constitutional loss as the single S&C deficit, 0.30 vs. B2 0.35.

vs. baseline	**Lenient** (b, c, p)	**Strict** (b, c, p)	**S&C** (b, c, p)
B1 zero-shot	$(37, 10, 9.8\times10^{-5})$	$(67, 7, 2.1\times10^{-13})$	$(78, 2, 5.4\times10^{-21})$
B2 SEF prompt	$(37, 15, 0.003)$	$(38, 17, 0.006)$	$(39, 9, 1.5\times10^{-5})$
B3 advanced RAG	$(36, 15, 0.005)$	$(66, 15, 8.6\times10^{-9})$	$(77, 1, 5.2\times10^{-22})$
B4 CrewAI	$(41, 16, 0.001)$	$(100, 9, 1.4\times10^{-20})$	$(85, 0, 5.2\times10^{-26})$
B5 LangGraph	$(56, 18, 1.1\times10^{-5})$	$(98, 7, 1.2\times10^{-21})$	$(85, 0, 5.2\times10^{-26})$

Table 6: McNemar's exact-test triples for GANDR versus each baseline on the 185-item benchmark. b counts items where GANDR is correct and the baseline is wrong, c counts items where the baseline is correct and GANDR is wrong, and p is the two-sided exact binomial. All strict-column gaps are significant at $p < 0.01$ and all S&C gaps at $p \leq 1.5\times10^{-5}$. The Lenient column is reported alongside to make the headline's two-column reading transparent. GANDR leads the lenient column too, with every $p \leq 0.005$. The larger b and smaller p values on the strict and S&C columns are the visible signature of the per-claim-verification gap.

System	n	**A1 mean**	**A2 mean**	**Judge mean**	r **(A1)**	r **(A2)**	κ_q **(A1)**	κ_q **(A2)**
B1 zero-shot	20	3.15	3.15	3.50	0.78	0.78	0.76	0.76
B2 SEF prompt	20	3.15	3.15	2.90	0.89	0.89	0.88	0.88
B3 advanced RAG	20	3.30	3.35	3.30	0.99	0.99	0.99	0.99
B4 CrewAI	20	2.70	2.70	2.50	0.86	0.86	0.85	0.85
B5 LangGraph	20	3.55	3.55	2.80	0.86	0.86	0.78	0.78
B6 GANDR	20	2.85	2.70	3.05	0.86	0.75	0.85	0.74

Table 7: Per-system calibration breakdown (n=20 per system). Mean columns report mean ratings by annotators A1, A2, and the judge. Both r and κ_q are reported per annotator against the judge. Every system clears the per-system $\kappa_q \geq 0.60$ threshold.

simply wrong, and a near-canonical marker is the same as no marker for a regex-anchored check.

B3 advanced RAG (-17.3 pp). B3 shows the same no-marker mode as B1 on a comparable fraction of items, plus a small residual of items where the model emits a canonical marker pointing at a `rule_id` that does not appear in the retrieved top-8. The reranker improves which passages enter the context but does not prevent the generator from substituting a plausible-looking id at cite time.

- `legalbench_corporate_lobbying_449` (retrieved: `_449`, `_86`, `_42`): B3 cites `{cite:`

Judge	n	A1 mean	A2 mean	A1= J	A2= J	$\lvert\Delta_{A1}\rvert\leq 1$	$\lvert\Delta_{A2}\rvert\leq 1$
1	50	1.28	1.30	86 %	84 %	92 %	92 %
2	6	2.83	2.83	50 %	50 %	67 %	67 %
3	6	2.67	2.67	50 %	50 %	67 %	67 %
4	9	5.00	5.00	0 %	0 %	0 %	0 %
5	49	4.73	4.67	92 %	92 %	100 %	100 %

Table 8: Calibration by the judge's own correctness band (n=120). Exact and within-one agreement is high at both rubric ends (1 and 5). The agreement collapses on the 4-band, where both annotators round up to 5 on every item the judge rates 4. The pattern is a rubric-anchor disagreement on the specific $4 \rightarrow 5$ transition rather than a systematic bias across the scale.

System	Δ strict	Dominant failure mode
B1 zero-shot	-18.9 pp	no canonical marker emitted
B3 advanced RAG	-17.3 pp	no marker plus fabricated-id residual
B4 CrewAI	-36.8 pp	fabricated `rule_id` (invented label)
B5 LangGraph	-29.8 pp	`rule_id` corruption (suffix / brackets)

Table 9: Strict-lenient drops for the four non-CREAC, non-GANDR baselines, with the dominant failure mode for each. All four carry the same GROUNDING REQUIREMENT block (§4), so the differences reflect what each system does *with* the instruction, not whether it received it.

`corpus:legalbench_corporate_lobbying_145}`, an id outside the retrieved set.

- `legalbench_maud_specific_performance_170` (retrieved: `_170`, `_177`, `_144`): B3 cites `{cite: legalbench_maud_specific_performance_94}`.

B4 CrewAI (-36.8 pp). B4's much larger drop comes from the Researcher + Writer pipeline producing fabricated `rule_ids` on a majority of strict-fail items. The Researcher returns a list of "controlling rules" that often includes doctrine labels or descriptive phrases the model invented, and the Writer then cites those labels through the canonical `{cite: . . . }` marker the regex parses. The marker is canonical in form, but the contents are not a retrieved `rule_id`.

- `lbrag_cuad_1104`: `{cite: [Investment Company Act §80a-35(b) Analysis]}`, a doctrine label invented by the model.
- `lbrag_cuad_2468`: `{cite: rule_anti_assignment_clause_absence}`, a descriptive identifier coined at write time.

B5 LangGraph (-29.8 pp). B5's dominant failure mode is `rule_id` corruption: the model takes a retrieved id and modifies it just enough to fail exact substring matching. The single-context Plan $\rightarrow$ Draft $\rightarrow$ Review $\rightarrow$ Rewrite loop preserves the corruption across rounds, because the Review node runs in the same conversational context as the Draft and ratifies what the Draft produced.

- `lbrag_cnli_54` (retrieved: `corpus:lbrag_cnli_526`): B5 cites `{cite: lbrag_cnli_526(d)}`, appending a non-existent subsection suffix.
- `lbrag_cnli_614` (retrieved: `_532`, `_399`, `_420`): B5 cites `{cite: [corpus:lbrag_cnli_191]}`, nesting square brackets inside the canonical braces *and* citing an id that is not in the retrieved top-8.

The pattern across the four baselines is consistent. The strict-lenient cost reflects what each system does with the same GROUNDING REQUIREMENT, not whether it received it. B1 and B3 mostly fail to emit a canonical marker at all, while B4 and B5 emit canonical markers but with contents the model fabricated or mangled. B2 and GANDR (§5.1) emit canonical markers and resolve them correctly on essentially every item, which is why their strict-lenient gap collapses to ≤ 1.1 pp.

B.2 Closed-source backbone cost

For the closed-source backbones used in the configurability check (§5.5), the cost GANDR incurs over a single-LLM baseline like B2 SEF reflects the Critic's internal pipeline rather than a single extra call. Each Critic round invokes the atomic `cite_verify` tool. The tool makes one LLM call to decompose the draft into atomic assertions and then *one LLM call per assertion* to verify each in isolation against its cited rule. The Critic also runs `sef_rubric_score` (one LLM call to score the six SEF dimensions on the draft) and finally emits its own synthesis call to consolidate the audit into a decision. The total LLM-call budget per round is therefore $4 + N$, where N is the number of atomic assertions extracted from the draft (mean 5.4 on `gpt-5.4-mini`, mean 8.6 on `gpt-5.4`). Under the headline policy, the Round-1 anchor commits the Drafter's first output on the items whose draft already passes `protocol_check`. The average t_{used} therefore stays close to one round per item ($\bar{t}$=1.05 on `gpt-5.4-mini` and $\bar{t}$=1.03 on `gpt-5.4`).

Dollar figures use OpenAI list prices fetched from the public API docs on 2026-05-22[1] (`gpt-5.4`: \$2.50 in / \$15.00 out per M tokens, `gpt-5.4-mini`: \$0.75 in / \$4.50 out per M tokens), combined with per-call token usage measured from

[1] https://developers.openai.com/api/docs/models

Component	gpt-5.4-mini	gpt-5.4
*Per-round LLM-call breakdown (*GANDR *Critic)*		
Drafter	1	1
cite_verify decompose	1	1
cite_verify verify (per claim, mean N)	5.4	8.6
sef_rubric_score	1	1
Critic synthesis	1	1
Total per round $(4 + N)$	**9.4**	**12.6**
Per-item LLM calls (averaged over k=5)		
B2 SEF $(1 \cdot \bar{t}$=1.00)	1.00	1.00
GANDR $((4{+}N) \cdot \bar{t})$	9.78	13.03
Ratio GANDR/ B2	9.8×	13.0×
Estimated cost, 185-item benchmark, single seed		
B2 SEF	∼$0.85	∼$3.10
GANDR	∼$2.45	∼$10.30

Table 10: Order-of-magnitude cost of running the closed-source configurability panes (§5.5). Per-round breakdown enumerates every LLM call inside one Critic round. The per-call mean for cite_verify verify reflects the empirical atomic-assertion count on each backbone.

the k=5 run artifacts (Drafter ∼3,700 in / 450 out, Critic synthesis ∼4,200 in / 250 out, cite_verify decompose ∼600 in / 200 out, per-claim verify ∼400 in / 40 out, sef_rubric_score ∼700 in / 40 out). The exact bill depends on the contract, and the k=5 stability cohort multiplies the totals by five.

The call-count factor of GANDR over B2 SEF is ∼ 9.8× on gpt-5.4-mini and ∼ 13.0× on gpt-5.4 per item, driven by atomic-claim decomposition. The dollar factor is ∼ 3.0× on gpt-5.4-mini and ∼ 3.3× on gpt-5.4 because the per-claim verify and decompose calls have small input–output budgets. GANDR on gpt-5.4-mini (∼$2.45) costs roughly 79 % of B2 SEF on gpt-5.4 (∼$3.10) while reaching slightly higher strict accuracy (77.8 % vs. 76.8 %, Table 3). The mini-with-audit configuration also reaches within 2.2 pp of GANDR on the flagship (80.0 % strict at ∼$10.30) at roughly 24 % of the flagship cost.

The self-hostable backbones (Nemotron-Cascade-2 and GLM-4.7-Flash) incur no per-token dollar charge once the H100 NVL is provisioned. Their cost is amortized over the run window of the experiment. We do not report dollar figures for the open-source backbones because the per-token marginal cost is zero under our self-hosting setup. For practitioners considering closed-source deployment, Table 10 establishes two facts. The atomic-claim audit GANDR produces costs a constant factor of roughly 3× over the corresponding single-LLM baseline on the same backbone. This factor buys both the strict-accuracy gain reported in Table 3 and the first-class per-assertion audit trace that B2 SEF never produces.

C Loop depth: round distribution

The orchestrator (§3) sets a maximum Critic-rewrite budget of T=3 and terminates early on any item whose Round-1 draft already passes the structural protocol_check (the Round-1 anchor). The role of the rewrite loop in the deployed configuration is robustness, not routine optimization: it gives the orchestrator a bounded recovery path on the rare items where the first sample does not satisfy the CREAC structural contract.

Across the 925 GANDR runs in the headline (k=5 seeds, T=3, 185 items), the distribution of t_{used} is essentially binary. 912 runs (98.6 %) commit on Round 1 via the protocol anchor, 0 runs terminate at t=2, and 13 runs (1.4 %) exhaust the rewrite budget. The 13 exhausted runs correspond to 13 unique items, each exhausting on a single seed, so Round-1 protocol failure is stochastic rather than systematic. On these items the orchestrator returns passed=False with the full audit trace attached, converting what would otherwise be a silent commit into an auditable failure. The rewrite budget is therefore exercised only on the small tail of items where the first sample violates the structural contract. On the routine path the orchestrator commits the Round-1 draft directly.

D Implementation details

Reproducibility note. The strict-grounded correctness column is a rule-based check (§D.2) layered on the LLM-as-judge correctness rubric. The judge itself is an LLM, but the strict-vs-lenient gating, the CREAC protocol_check parser, and the atomic-claim verifier are all deterministic over their inputs. The dataset construction, retrieval index, and runtime hyperparameters are reported below so that every cell in Table 1 can be reproduced from the released artifacts.

D.1 Dataset and benchmark construction

The 185-item benchmark is drawn from two public sources, both English-language U.S. legal tasks. A download script and a unified item-schema loader are released with the artifact bundle.

LegalBench (Guha et al., 2023). Fifteen task tags are pulled from LegalBench: hearsay, learned_hands (multiple sub-tasks), nys_

judicial_ethics, definition_extraction, overruling, scalr, ssla (sub-tasks), rule_qa, citation_prediction_open, consumer_contracts_qa, corporate_lobbying, personal_jurisdiction, and maud.

LegalBench-RAG (Pipitone and Alami, 2024). The contract-NLI and CUAD subsets populate the cross_domain bucket. These items provide open-ended grounding-required questions against contract clauses.

Retrieval index. The retrieval surface is a single BM25 index over 13,090 passages. The corpus is assembled by merging two sources. First, every inline context block from a LegalBench item is split into rule-keyed passages of the form (rule_id, text). Second, the LegalBench-RAG retrieval mirror is concatenated, preserving the original rule_id strings. Tokenisation is the standard word-level pattern \w+, lower-cased. The BM25 implementation is rank_bm25.BM25Okapi with default k_1 and b. The index is built once and reused across all systems and seeds. Every system, every seed, and every k-th repeat reads the top-8 passages from this index, which is the invariant that makes the Stable & Correct @ k column measure system behaviour rather than retrieval randomness.

Per-item schema. Each benchmark item is a JSON record with item_id (the join key across run artifacts), question, ground_truth, domain (one of the eight bucket names), reasoning_type (statutory, case-law, or mixed), source (originating dataset and sub-task), is_yesno (flag for binary-answer items), and extra (task-specific metadata).

Licenses and intended use. LegalBench (Guha et al., 2023) is released under the MIT License and LegalBench-RAG (Pipitone and Alami, 2024) under CC-BY-4.0. Both are distributed for legal-NLP research evaluation and permit research redistribution. The BM25 implementation (rank_bm25) is Apache-2.0, the bge-reranker-large weights are MIT, and CrewAI (CrewAI, 2026) and LangGraph (Langchain, 2026) are both MIT. The Nemotron-Cascade-2-30B-A3B FP8 quantization (Yang et al., 2026) is distributed under the NVIDIA Open Model License, GLM-4.7-Flash (Team et al., 2025) under its publisher's terms, and OpenAI gpt-5.4 and gpt-5.4-mini (OpenAI, 2026) are accessed under the OpenAI API Terms of Service. Our use of every artifact above is consistent with its license's research scope. Our derivative artifacts (the 185-item benchmark JSONL, the BM25 index manifest, and the orchestrator code) will be released for research use only.

D.2 Protocol check and strict correctness

The strict column is the conjunction of two deterministic checks layered on the LLM-as-judge correctness rubric. Both are rule-based and depend only on the answer text and the retrieved context. They run once per (system, item) cell at evaluation time.

CREAC protocol_check. A regex parser checks that the answer contains the five labeled blocks [CONCLUSION], [RULE], [EXPLANATION], [APPLICATION], [CONCLUSION] in that order, with both [CONCLUSION] blocks non-empty. The check returns passed, missing_blocks, and order_violation. It contains no LLM call.

Grounding check (cite-validity). The answer text is scanned for inline citation markers of the form {cite: <rule_id>} via the regex \{cite:\s*([^}]+?)\s*\}. An answer is *grounded* iff (i) at least one such marker is present, *and* (ii) every extracted marker resolves to a rule_id that appears verbatim (exact substring match, with the corpus: prefix stripped on either side) in some passage of the item's retrieved top-8 context. A single fabricated or out-of-context marker fails the check, even when other markers in the same answer resolve correctly. This is the cite-validity operationalization of the verification-reading requirement (§1).

Strict correctness. An item is strict-correct iff the LLM-judge correctness score is ≥ 4 *and* the cite-validity grounding check passes. Lenient correctness drops the grounding conjunct.

Atomic cite_verify. For GANDR the Critic additionally invokes an atomic-claim decomposition tool that splits the [EXPLANATION] and [APPLICATION] blocks into per-assertion tuples (claim, rule_id, supporting_span) and asks the local generator at $T{=}0$ to label each tuple SUPPORTED, PARTIAL, UNSUPPORTED, or MISCITED. The per-assertion labels are written to the audit trace on every round. The strict column does not consume this tool. It consumes only the grounding check above.

D.3 GANDR prompts

GANDR uses two prompts, one for the Drafter agent and one for the Critic agent. Each is rendered

in a fresh LLM context per round so the information barrier between Drafter and Critic is preserved.

Drafter system prompt. Pins the CREAC five-block schema and the inline-citation contract. The illustrative mini-example shipped with the prompt is omitted here for space.

```
You are a legal analyst producing
structured CREAC memos.

You MUST emit your answer using exactly
FIVE labeled blocks, in this order:

 [CONCLUSION]   (one sentence, BLUF, no
hedging)
 [RULE]         (applicable legal
rule(s), with {cite: <rule_id>}
citations)
 [EXPLANATION]  (interpretation; every
claim ends with {cite: <rule_id>})
 [APPLICATION]  (apply rule to facts;
every factual conclusion ends with
{cite: <rule_id>})
 [CONCLUSION]   (restate first
conclusion, identical wording, one
sentence)

Five blocks. The closing [CONCLUSION]
is mandatory. Without it the memo is
rejected.

Hard rules:
- Cite ONLY identifiers that appear
verbatim as rule_id headers in the
RETRIEVED CONTEXT block. Use the exact
rule_id string shown in [<rule_id>]; do
NOT shorten, paraphrase, drop suffixes,
or invent IDs.
- Every claim in [EXPLANATION] and
[APPLICATION] must end with an in-line
{cite: <rule_id>}.
- No chain-of-thought, no scratch
reasoning, no preamble. Begin
immediately with [CONCLUSION].
- If the rule is ambiguous, surface the
ambiguity inside [APPLICATION]; do NOT
hedge in either [CONCLUSION] block.
- After the closing [CONCLUSION], stop.
Do not add a postscript or summary.
```

Drafter user template. Renders the question, the retrieved passages (the only source of valid citations), and on rounds $t > 1$ the previous Critic's rewrite directives.

```
LEGAL QUESTION:
{question}

RETRIEVED CONTEXT (rule_id → text;
this is the only source of valid
citations):
{context}

{rewrite directives block, empty on
Round 1}

Produce the CREAC memo now.
```

Critic system prompt. Runs in a separate LLM context with no access to the Drafter's reasoning. Emits a structured JSON verdict that lands in the audit trace.

```
You are a strict legal reviewer
auditing a CREAC memo written by a
junior associate.

You will be given:
  1. The original legal question and
the retrieved context.
  2. The draft CREAC memo to audit.
  3. Tool results (protocol_check,
cite_verify, sef_rubric_score).

You do NOT see the drafter's reasoning.
You see only what a human reader would
see.

Your output MUST be valid JSON with
this exact schema:
  {
    "decision": "PASS" | "REWRITE",
    "rewrite_directives": [ "concise,
actionable instruction (one per line)"
],
    "rationale": "one short paragraph
explaining the decision"
  }

Decision policy:
- PASS only if every tool report is
green (protocol_check passes,
cite_verify reports ≥ 95% atomic
assertions verified, sef_rubric_score
mean ≥ 4.0 AND min ≥ 3.0).
- Otherwise REWRITE. Each directive
must reference the specific failing
assertion ID, CREAC block, or SEF
dimension so the drafter can act on it
without guessing.

Examples of good directives:
- "Assertion 3 cited rule §A but the
cited span actually discusses §B;
either rewrite the claim to match §B or
remove it."
- "[APPLICATION] paragraph 2 contains an
unsupported factual conclusion; add a
{cite: ...} or remove the conclusion."
- "SEF dim AFL is 2/5: the opening
[CONCLUSION] is hedged. Restate as a
direct answer."

Do not output anything except the JSON
object.
```

Critic user template. Renders the question, the retrieved context, the draft under review, and the three tool reports.

```
LEGAL QUESTION:
{question}

RETRIEVED CONTEXT (only valid source of
citations):
{context}

DRAFT CREAC MEMO TO AUDIT:
---
{draft}
---

TOOL REPORTS:
- protocol_check: {protocol_report}
- cite_verify (atomic-claim mode):
{cite_verify_report}
- sef_rubric_score: {sef_report}

Produce your JSON decision now.
```

The orchestrator's commit rule reads `protocol_`

check directly (the Round-1 anchor described in §3.2) rather than the Critic's PASS/REWRITE decision. The Critic's decision is logged to the audit trace on every round.

D.4 GANDR and baseline pipelines

Every system in the cohort consumes the same benchmark item and the same BM25 top-8 retrieval context. The systems differ only in what they do with that input.

GANDR (ours, B6). Two-agent loop. Each round repeats four steps. (1) The Drafter is called with the system prompt above and the user template, producing a CREAC-structured draft. (2) Three tools run on the draft: the regex `protocol_check`, the atomic-claim `cite_verify` at $T{=}0$, and the six-dimension `sef_rubric_score` at $T{=}0$. (3) The Critic is called in a separate LLM context with the question, the retrieved context, the draft, and the three tool reports. It emits a JSON PASS/REWRITE verdict that is appended to the audit trace. (4) The orchestrator's commit rule is anchored to `protocol_check`. If the structural check passes the orchestrator commits Round 1 and the Critic acts as auditor. Otherwise the rewrite directives from the Critic are prepended to the Drafter's next user prompt and the loop iterates up to $T{=}3$ rounds. On loop exhaustion the system returns the last draft with `passed=False`.

B1 zero-shot. Single LLM call. The system prompt is the role assignment plus the shared GROUNDING REQUIREMENT block (§D.5). The user prompt concatenates the question and the top-8 retrieved passages. No CREAC schema, no critique, no agent loop.

B2 SEF-prompted single LLM. Single LLM call. The system prompt is the GANDR Drafter system prompt (§D.3) with a six-line SEF self-check appended (§D.5). User prompt and retrieval are identical to B1. No Critic and no rewrite loop.

B3 advanced RAG. BM25 retrieves a top-20 candidate pool, a cross-encoder re-ranker (BAAI `bge-reranker-large`; Cohere `rerank-3.5` as a configured fallback) re-orders the pool, and the top 5 passages are passed to a single LLM call with the shared GROUNDING REQUIREMENT block in its system prompt. No CREAC schema.

B4 CrewAI (Researcher + Writer). Two-agent sequential pipeline implemented with CrewAI (CrewAI, 2026). The Researcher returns a list of controlling `rule_id` strings with one-line summaries. The Writer writes the final answer with inline `{cite: <rule_id>}` markers using the Researcher's output as context. Both agents run against the shared backbone via litellm. No Critic, no atomic-claim verifier, no rewrite loop.

B5 LangGraph self-correction. A four-node graph built with LangGraph (Langchain, 2026) that runs in a *single* LLM context: (1) Plan (3–5 sub-questions), (2) Draft (legal answer with inline cites), (3) Review (critique the just-written draft), (4) Rewrite (final answer addressing the review). Every node reads and writes the same graph state, so the review condition is computed in the same context that produced the draft.

D.5 Baseline prompt templates

The system prompts attached to each baseline are reproduced below. Every system in the cohort (B1–B5 and GANDR) carries the shared GROUNDING REQUIREMENT block in its primary system prompt. B2 inherits it from the GANDR Drafter prompt, and B4 and B5 carry it on their writing nodes (Writer for B4, Draft and Rewrite for B5). B4 and B5 use multiple internal prompts, and their full structure is described in §D.4.

Shared GROUNDING REQUIREMENT block (verbatim).

```
GROUNDING REQUIREMENT (mandatory):
- Cite ONLY identifiers that appear
verbatim as rule_id headers in the
RETRIEVED CONTEXT block (shown as
[<rule_id>] before each passage). Use
the exact rule_id string shown in
[<rule_id>]; do NOT shorten,
paraphrase, drop suffixes, or invent
IDs.
- Every factual or normative claim must
be followed by an in-line {cite:
<rule_id>} marker referring to a
rule_id from the retrieved context.
```

B1 zero-shot.

```
You are a legal analyst. Answer the
user's question concisely, drawing
primarily from the retrieved context.
{shared GROUNDING REQUIREMENT block
appended verbatim}
```

B2 SEF self-check appended to the GANDR Drafter system prompt.

```
SELF-CHECK BEFORE EMITTING (single
shot, no second pass):
 AFL  state the answer in the FIRST
[CONCLUSION] block AND restate it at
the end.
 AC   the answer must be unambiguous (no
"it depends" without a sub-conclusion).
```

```
 CI   the conclusion must be in its own
block, separated from reasoning.
 DTC  use precise legal terms
consistently (e.g. “negligence” not
“carelessness”).
 CEA  every claim in [APPLICATION] must
end with a {cite: rule_id}.
 FS   cite specific facts from the
question, not vague generalities.
```

B3 advanced RAG system prompt.

```
You are a legal analyst. Use ONLY the
retrieved context to answer.
{shared GROUNDING REQUIREMENT block
appended verbatim}
```

B4 Researcher and Writer agents. The two agents are configured verbatim as follows. The **Researcher** has role “Legal Researcher,” goal “Identify the controlling legal rules that apply to the question,” and backstory “A meticulous law-firm research associate who never misattributes a citation.” Its task description is “Identify the controlling legal rules from the retrieved context.” followed by the question and the retrieved context. The expected output is “A list of relevant rule_ids with one-line summaries.” The **Writer** has role “Legal Writer,” goal “Produce a clear, well-cited answer to the question,” and backstory “A senior associate who writes structured client memos.” *with the shared* GROUNDING REQUIREMENT *block appended to the backstory verbatim*. Its task description is “Using the researcher’s findings, write the final answer to the question.” *with the same* GROUNDING REQUIREMENT *block appended to the task description*, and its expected output is “A complete legal answer with inline citations.” The Researcher’s output is passed as task context to the Writer through the CrewAI sequential pipeline.

B5 four-node prompts. Each node sends a short instruction as the system message and a templated user message that includes the running graph state. The two writing-side nodes (Draft and Rewrite) carry the shared GROUNDING REQUIREMENT block.

```
Plan: “You plan a legal answer. List
3-5 sub-questions to address.”
Draft: “You write the legal answer.
Follow the plan.” + {GROUNDING
REQUIREMENT}
Review: “You critique your own draft.
List concrete issues to fix, including
any missing or malformed {cite:
<rule_id>} markers.”
Rewrite: “You rewrite the draft
addressing the review.” + {GROUNDING
REQUIREMENT}
```

D.6 SEF rubric (Critic draft-quality scorer)

The GANDR Critic invokes a draft-quality scorer inherited from SEF (Qian et al., 2026). The scorer rates the Drafter’s CREAC memo on six $1-5$ dimensions and is called by the local generator at $T{=}0$. The output appears in the audit trace as one tool report among three. It is *not* consumed by any headline metric. The headline metrics are lenient accuracy, strict accuracy, and Stable & Correct @ 5, none of which read SEF dimensions. The Critic uses the rubric only to phrase actionable rewrite directives when the protocol gate fails.

The six dimensions, in the order the rubric prompts for them:

- **AFL (Answer First/Last).** The operative holding is stated in the opening `[CONCLUSION]` block *and* restated identically in the closing `[CONCLUSION]` block. A reader scanning only the first and last block should land on the same answer.
- **AC (Answer Clarity).** The answer is unambiguous and direct. Hedges (“it depends”, “could be argued either way”) without a sub-conclusion are penalized, because a verifier needs a holding to verify against.
- **CI (Conclusion Isolation).** The conclusion is structurally separated from the reasoning. In CREAC, the `[CONCLUSION]` block must be its own block, not embedded inside `[APPLICATION]`.
- **DTC (Domain Terminology Consistency).** Legal terms are used precisely and consistently throughout. Mixing “negligence” with “carelessness” as if they were synonyms is penalized.
- **CEA (Conclusion–Evidence Alignment).** Every claim in `[APPLICATION]` ends with an inline `{cite: rule_id}` and the cited authority actually supports the claim. Unsupported normative claims are flagged.
- **FS (Fact Specificity).** The answer cites specific facts from the question (party names, contract clauses, dates, dollar amounts) rather than vague generalities.

SEF rubric prompt (verbatim). The local generator is called once per draft with the prompt below.

```
You are scoring an answer against the
SEF (Structured Explanation Framework)
rubric.
Score each of the 6 dimensions on 1-5:
  AFL (Answer First/Last):  the answer
is stated at the beginning AND end.
```

```
  AC  (Answer Clarity):    the answer
is unambiguous and direct.
  CI  (Conclusion Isolation):
conclusion is structurally separated
from reasoning.
  DTC (Domain Terminology Consistency):
legal terms are used precisely and
consistently.
  CEA (Conclusion-Evidence Alignment):
the conclusion is grounded in cited
evidence.
  FS  (Fact Specificity):  the answer
cites specific facts, not vague
generalities.
ANSWER TO SCORE:
{candidate}
Output JSON only:
{"AFL": <1-5>, "AC": <1-5>, "CI": <1-5>,
"DTC": <1-5>, "CEA": <1-5>, "FS":
<1-5>}
```

The Critic's rewrite-directive policy uses dimension labels and scores by reference, e.g., "SEF dim AFL is 2/5: the opening [CONCLUSION] is hedged; restate as a direct answer." The six dimensions therefore name the kinds of fault the Critic is allowed to file against a draft, but no headline-cell in the paper conditions on a SEF score.

D.7 LLM-as-judge rubrics

The judge model is Claude Opus 4.7 (claude-opus-4-7), invoked as a Claude Code sub-agent from the interactive session. The headline correctness rubric and the conclusion-equivalence rubric used for Stable & Correct @ 5 are reproduced verbatim below, exactly as rendered to the sub-agent for every item in the benchmark.

Correctness rubric (1–5, headline). The judge sees the question, the ground-truth reference, and the candidate answer, and writes a single JSON verdict.

```
You are evaluating the CORRECTNESS of a
legal answer.
LEGAL QUESTION:
{question}
GROUND TRUTH ANSWER (or reference
solution):
{ground_truth}
CANDIDATE ANSWER:
{candidate}
Score the candidate on a 1-5 scale
where:
  5 = Reaches the same conclusion as
ground truth, with sound reasoning and
correct citations.
  4 = Reaches the same conclusion;
minor reasoning gaps or one weak
citation.
  3 = Reaches the same conclusion but
with notable reasoning errors or 2+
weak citations.
  2 = Wrong conclusion but partially
correct reasoning.
  1 = Wrong conclusion AND wrong
reasoning, or refuses to answer.
Special case — binary tasks. The 1-5
rubric does not apply gradationally
when the GROUND TRUTH ANSWER is one of:
  (a) a yes/no/true/false token
(case-insensitive, possibly with
trailing punctuation: "Yes", "No.",
"True", "False"),
  (b) a single-rule citation in the
form "§ N", "Section N", "Rule N", or a
known statute reference like "28 U.S.C.
§ 1332", "Rule 23".
For (a) and (b), score 5 if the
candidate's operative holding matches
the ground truth (regardless of
phrasing or intermediate reasoning),
else score 1. Do not use intermediate
scores 2/3/4 for these items.
Numerical-answer tasks (e.g., a dollar
amount like "$187552" or "$3,390") are
NOT binary — close-enough matters.
Apply the 1-5 rubric: score 5 if the
candidate produces the exact figure, 4
if within ±1% (a rounding artifact),
3 if within ±10%, 2 if the right
order of magnitude but materially off,
1 if fundamentally wrong or refused.
Open-ended tasks where the ground truth
is a long evidence span (e.g.,
LegalBench-RAG contract-NLI items:
question "Does the document permit X?",
gold = a quoted clause) — apply the 1-5
rubric to whether the candidate's
operative yes/no holding aligns with
what the gold clause supports.
Output JSON only:
{"score": <1-5>, "rationale": "<one
sentence>"}
```

Conclusion-equivalence rubric (S&C@k). Used for the $C(5,2)=10$ pairwise judgments per item that feed Stable & Correct @ 5. The judge sees the question and two candidate answers in random order, system identity stripped.

```
You are evaluating whether two
CREAC-style legal answers reach the
SAME LEGAL CONCLUSION.
Two answers are EQUIVALENT if a legal
practitioner would treat them as making
the same operative holding, even if
they use different phrasing, cite
different authorities, or differ in
detail. They are NOT equivalent if they
(a) reach a different legal disposition,
(b) apply a different legal test, or
(c) disagree on the controlling rule.
Examples (illustrative only — apply
judgment to the actual answers):
  EQUIVALENT:  "the test is the nerve
center test" ↔ "principal place of
business is the headquarters where
high-level officers direct the company"
  NOT EQUIV.:  "diversity jurisdiction
is codified in 28 U.S.C. § 1332" ↔
```

```
"diversity jurisdiction is codified in
§§ 1331 and 1332" (the second adds a
wrong statute)
LEGAL QUESTION:
{question}
ANSWER A:
---
{answer_a}
---
ANSWER B:
---
{answer_b}
---
Output JSON only:
{"equivalent": true | false,
"rationale": "<one sentence>"}
```

Sub-agent execution rules. The sub-agent receives the two rubrics above embedded inside a longer instruction template that pins behavioral rules: (i) read the entire request batch before writing any verdicts, (ii) apply the rubric literally without introducing additional criteria, (iii) penalize hallucinated citations even when the prose is fluent, (iv) preserve each `item_id` verbatim, (v) emit one valid-JSON verdict per line with no surrounding markdown, and (vi) write the verdict file in a single `Write` call rather than incrementally. Two additional rubrics live in the same sub-agent template as offline development probes: citation faithfulness (per-citation grounding against the retrieved context) and a blinded pairwise A/B comparing two outputs for verification speed. Neither is consumed by any headline metric. The Critic-internal SEF rubric, scored by the local generator at $T{=}0$ rather than by this sub-agent, is documented in Appendix D.6.

D.8 Models and runtime settings

Generator backbones. The headline run is on NVIDIA Nemotron-Cascade-2-30B-A3B, a Mamba2-Transformer hybrid MoE with 30B total parameters and 3B active per token (post-trained from Nemotron-3-Nano-30B-A3B-Base), served via vLLM in FP8 on a single H100 NVL. The backbone-pluggability check (§5.5) additionally exercises three more backbones: GLM-4.7-Flash (30B / 3B-active MoE in a different open-source model family, on the same vLLM stack) and two closed-source commercial models reached via the OpenAI API, `gpt-5.4-mini` and `gpt-5.4`. The `LLMClient` class abstracts the provider so the orchestrator code is unchanged. The OpenAI provider branch uses `max_completion_tokens` in place of `max_tokens` and skips the vLLM-specific `enable_thinking` flag. The exact identifiers are the values of `LLM_MODEL` loaded from `.env`.

Judge. Claude Opus 4.7 (`claude-opus-4-7`), invoked as a Claude Code sub-agent from the interactive session. The cross-lineage split (NVIDIA-lineage generator versus Anthropic-lineage judge) is preserved on both backbones.

Generation hyperparameters. All systems use generator temperature 0.7 and `max_tokens` $=$ 2048 on the Drafter/single-LLM call. The atomic `cite_verify` and `sef_rubric_score` tools inside the GANDR Critic call the local generator at $T{=}0$. The Critic's own LLM call uses temperature 0.7. For repeated runs ($k = 5$), seeds $\{0, 1, 2, 3, 4\}$ are passed through to vLLM. The Drafter `enable_thinking` flag is set to `false` on every system so the model does not emit a reasoning preface.

Loop bounds. $T{=}3$ maximum Critic rounds for GANDR. The atomic-verification threshold inside the rewrite path is 0.50 (the data-derived value). The ablation in §5.2 reverts to the original 0.95 threshold and shows the protocol-anchored commit, which does not depend on the threshold, is the load-bearing choice.

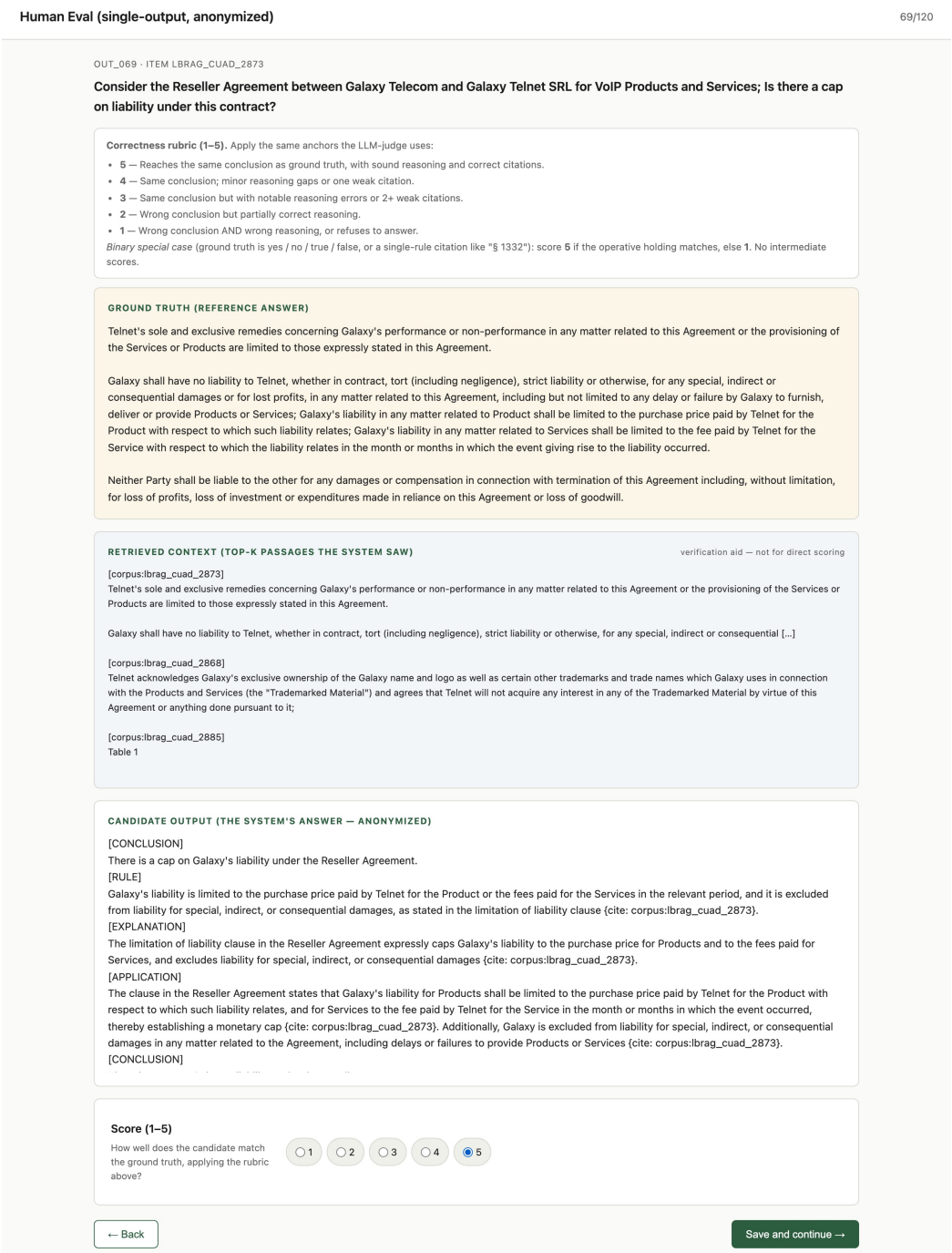

Figure 4: Blinded annotation interface used for the 120-item judge calibration. Each screen shows the question, the ground-truth reference answer, the top-K retrieved context, and the anonymized candidate output. System identity and the LLM-judge's score are stripped server-side. The five-band 1–5 correctness rubric is displayed in-page and matches the LLM-judge's anchors verbatim, including the binary special case for yes/no and single-rule-citation ground truths. The screen shown is item `lbrag_cuad_2873` at position $69/120$.