BEYOND SEMANTIC RELEVANCE: LEARNING DECISION AUTHORITY OVER RETRIEVED PRECEDENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models increasingly use retrieved examples and cases to support prediction, but semantic relevance alone does not determine whether a past decision should transfer to a new query. This limitation is especially important in case-augmented reasoning, where retrieved examples carry historical decisions that can directly steer the current prediction. We formulate this problem as *decision-authority learning*: determining when a retrieved precedent is applicable enough for its prior decision to influence the current query. A key challenge is that outcome supervision cannot distinguish valid from invalid precedent transfer when both lead to the same correct label, a failure mode we call *analogical reward aliasing*. To address this, we construct *contrastive intervention groups* (CIGs) that approximately match retrieval relevance while varying applicability, and introduce *Grounded Analogical Reward Optimization* (GARO) together with its intervention-grouped variant, IG-GARO, to separate task correctness from precedent-transfer quality. A separate utility model controls when the trained specialist should replace Query-only. On natural similarity-matched COLD pairs, IG-GARO improves authority discrimination accuracy from 73.6% to 87.6%. The full system reaches 88.1 Macro-F1 while invoking the specialist on 19.7% of queries, and achieves the highest Macro-F1 point estimate across all evaluated moderation and legal settings.

1 INTRODUCTION

Large language models increasingly rely on retrieval to incorporate external knowledge, demonstrations, and prior cases into prediction. In ordinary retrieval-augmented generation, the retrieved item is mainly a source of information. Case-augmented reasoning is different: a retrieved case carries not only evidence or context, but also a historical decision. Once such a case is used as an analogy, its prior label or judgment can directly steer the current prediction. This makes case retrieval useful, but also creates a failure mode that semantic relevance alone cannot resolve.

Content moderation provides a concrete example. Two statements may mention the same target and share most of their wording while differing in stance, quotation, endorsement, or intent. In Figure 1, the closest retrieved case condemns an exclusionary statement, whereas the query endorses it. The retrieved case is highly relevant, yet transferring its historical decision produces the wrong prediction. Similar issues arise in legal precedent reasoning, where a seemingly close case may cease to apply once a material factual or procedural distinction is taken into account (Ashley, 1991). Retrieval can therefore identify plausible candidates, but it does not determine which retrieved decisions should be trusted. This distinction is not captured by standard task supervision. Suppose two precedent-conditioned trajectories both reach the correct label. One may rely on a genuinely applicable precedent, while the other uses an inapplicable case whose historical decision happens to agree with the target. From the final outcome alone, the two trajectories receive the same credit, even though only one reflects valid transfer. We refer to this ambiguity as *analogical reward aliasing*. It motivates a separate learning problem: *decision-authority learning*, which asks whether a retrieved precedent is applicable enough for its historical decision to influence the current prediction.

Our approach makes this transfer boundary explicit. We construct *contrastive intervention groups* (CIGs) for the same query, consisting of a no-precedent self-anchor, an applicable precedent, and a similarity-matched materially inapplicable precedent. By keeping retrieval relevance approximately

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

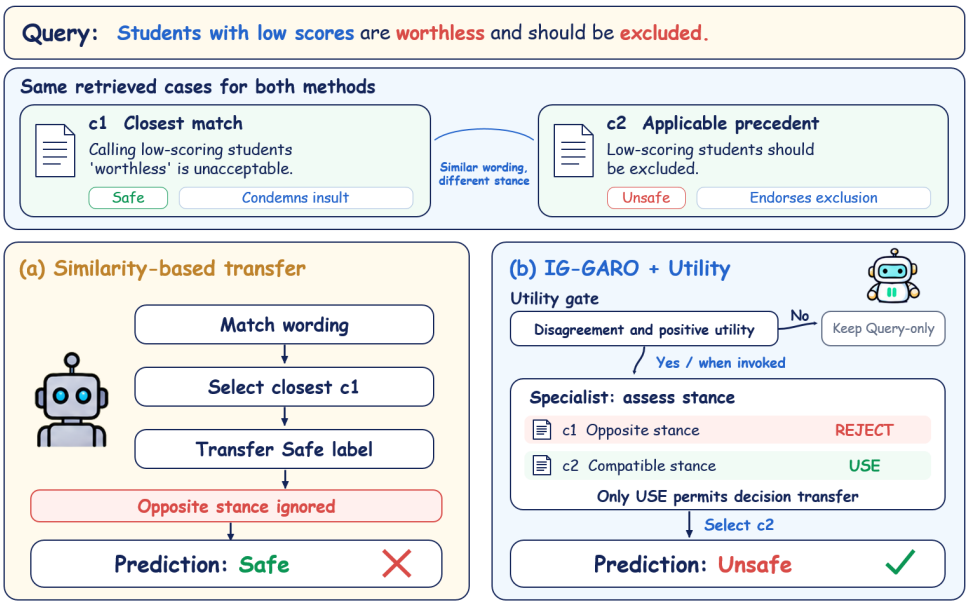


Figure 1: Why retrieval similarity is insufficient for precedent transfer. Similarity-based transfer can copy the decision of a highly similar but inapplicable case, whereas our system grants decision authority only to precedents predicted as USE.

matched while changing applicability, CIGs expose whether a model is responding to the decision-relevant relation rather than to topical similarity alone. The specialist predicts a structured transfer state—USE, LIMIT, or REJECT—and only USE candidates are allowed to transfer their historical decisions. To learn from these interventions, we introduce *Grounded Analogical Reward Optimization* (GARO), which separates task correctness from process-level feedback about authority and precedent relations. GARO evaluates whether the transfer state is appropriate and whether the supporting relation is grounded and material. Its intervention-grouped variant, IG-GARO, changes the optimization unit from an individual precedent to the full matched intervention group, so that valid and invalid transfers for the same query are compared directly. A pairwise authority objective further ranks applicable precedents above materially inapplicable ones. The result is a specialist that is trained not merely to use retrieved cases, but to decide which retrieved decisions deserve authority. Authority learning, however, does not imply that the specialist should be invoked on every query. Our experiments show that precedent-conditioned reasoning is strongly helpful in some regions of the input space but harmful in others: it can repair a Query-only error when retrieved cases disagree, yet overwrite an already correct prediction when the base model is confident. We therefore separate precedent authority from deployment utility. A lightweight utility estimator, trained from out-of-fold branch outcomes, predicts when invoking the specialist is preferable to retaining Query-only. Authority determines which precedent may influence the specialist; utility determines whether the specialist should influence the final system output.

Experiments on COLD isolate both effects. On a natural similarity-matched authority set, IG-GARO improves authority discrimination accuracy from 73.6% with Single-case GARO to 87.6%. With selective deployment, the full system reaches 88.1 Macro-F1 while invoking the specialist on only 19.7% of queries. The same formulation remains effective across additional Chinese moderation benchmarks, different model families and scales, and independently trained English legal tasks. These results suggest that the key challenge in case-augmented reasoning is not simply retrieving more similar examples, but learning which historical decisions are transferable and when using them is worthwhile.

In summary, this paper makes the following key contributions: First, we formulate *decision-authority learning* and distinguish retrieval relevance, precedent applicability, decision authority, and deployment utility in case-augmented reasoning. Second, we introduce CIGs, GARO, and IG-GARO to learn precedent transfer from matched interventions rather than task outcomes alone.

Third, we show that precedent-conditioned reasoning provides strong but conditional expertise, and that selective deployment is necessary to convert this local capability into consistent end-to-end gains across moderation and legal reasoning settings.

2 Related Work

Case-augmented reasoning and retrieval. Retrieval-augmented models use external information to support generation (Lewis et al., 2020), and semantically similar demonstrations and their labels can substantially affect in-context predictions (Liu et al., 2022; Min et al., 2022; Yoo et al., 2022). Class-RAG, CARO, and CHAIRO bring retrieval or analogy-based reasoning into content moderation (Chen et al., 2024; Wu et al., 2026; Lu et al., 2026). We study a complementary problem: retrieved cases can remain close in semantic space while differing in whether their historical decisions are transferable. Legal NLP provides a natural cross-domain test because precedent relevance and applicability are also distinct there (Ashley, 1991; Semo et al., 2022; Putta et al., 2026).

Process feedback and selective computation. Process supervision can distinguish trajectories that reach the same answer through different reasoning paths (Uesato et al., 2022; Lightman et al., 2023). GARO applies this idea to precedent transfer by evaluating whether the underlying relation is grounded and material, while GRPO supplies the group-relative optimization mechanism (Shao et al., 2024). Selective classification and metareasoning study when to abstain or allocate additional computation (Geifman & El-Yaniv, 2017; Russell & Wefald, 1991). Here, routing instead chooses between Query-only and a decision-producing precedent specialist.

3 Decision Authority in Precedent-Augmented Reasoning

A precedent-augmented pipeline contains four stages: retrieval proposes candidate cases, applicability determines whether a retrieved case can function as a precedent for the current query, decision authority determines whether its historical decision may affect the specialist, and deployment utility decides whether the specialist should replace Query-only.

3.1 Applicability and Specialist Composition

For query x, a fixed retriever returns $\mathcal{C}_k(x) = \{c_1, \ldots, c_k\}$ with labels y_{c_j} and scores s_j. The Query-only generalist predicts $f_Q(x)$. For each query–precedent pair, the specialist predicts

$$a_\theta(x, c) \in \{\text{USE}, \text{LIMIT}, \text{REJECT}\}. \tag{1}$$

USE means that the precedent's historical decision is sufficiently applicable and may influence the current prediction. LIMIT represents partial transfer: the precedent provides contextual evidence or constraints, but its historical decision should not be directly transferred. REJECT means that the retrieved case is not decision-relevant for the current query. We supervise all three states because LIMIT captures the boundary between useful evidence and transferable authority; however, only USE cases enter the authority set for decision transfer. Applicability is the structured relation judgment in Eq. (1); decision authority is the resulting permission to transfer the precedent's historical decision.

$$D_\theta(x, c) = \mathbb{I}[a_\theta(x, c) = \text{USE}], \qquad A_\theta(x, c) = \log p_\theta(\text{USE} \mid x, c) - \log p_\theta(\text{REJECT} \mid x, c), \tag{2}$$

where D_θ determines eligibility and A_θ ranks eligible precedents. Let $\mathcal{J}_{\text{auth}}(x) = \{j : D_\theta(x, c_j) = 1\}$ and

$$j^\star = \arg \max_{j \in \mathcal{J}_{\text{auth}}(x)} A_\theta(x, c_j). \tag{3}$$

The specialist uses $c_{j^\star}$ when the authority set is nonempty and otherwise falls back to the no-precedent self-anchor $f_\theta(x, \varnothing)$.

3.2 Deployment Utility

Authority is local to the specialist; deployment is separate. Let $r^\star(x)$ indicate that the specialist repairs a Query-only error and $b^\star(x)$ indicate that it overwrites a correct Query-only prediction. Given inference-time observables $v(x, \mathcal{C}_k)$, we learn

$$U_\phi(x, \mathcal{C}_k) = q_\phi(v(x, \mathcal{C}_k)) \approx \mathbb{E}[r^\star(x) - \lambda_b b^\star(x) - \eta \mid v(x, \mathcal{C}_k)]. \tag{4}$$

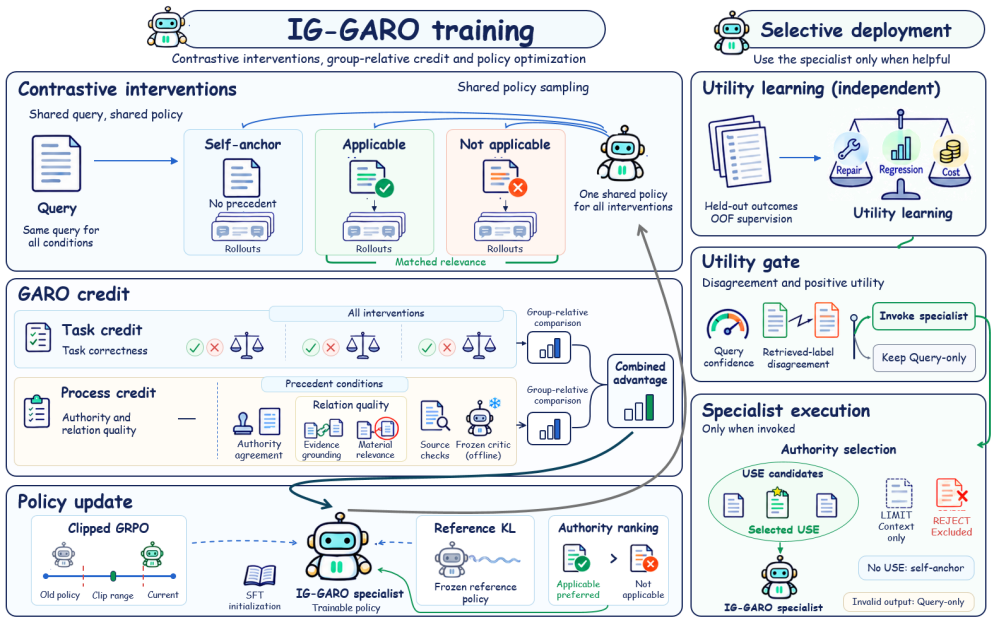


Figure 2: Training and selective deployment of the decision-authority specialist. IG-GARO trains a shared specialist over contrastive intervention groups containing a self-anchor, an applicable precedent, and a similarity-matched non-applicable precedent. GARO separates task credit from process credit, while intervention-grouped optimization and authority ranking shape precedent transfer. A separately trained, cross-fitted utility model then decides whether the specialist should replace Query-only; once invoked, only USE precedents are eligible for decision transfer.

4 LEARNING DECISION AUTHORITY

Figure 2 gives the full training and deployment pipeline: IG-GARO learns the specialist from contrastive intervention groups, while the utility model is trained separately from out-of-fold branch outcomes.

4.1 CONTRASTIVE INTERVENTION GROUPS

For each training query, we construct

$$\mathcal{I}_x = \{c^{(0)}, c^{(+)}, c^{(-)}\}, \qquad c^{(0)} = \varnothing, \quad \mathrm{Sim}(x, c^{(+)}) \approx \mathrm{Sim}(x, c^{(-)}), \tag{5}$$

where $c^{(+)}$ is fully applicable and $c^{(-)}$ is materially non-applicable. Matching retrieval relevance makes applicability the main source of variation. BGE-M3 retrieves candidate precedents, and an offline teacher extracts task-specific relation signatures used only for construction. Natural matched pairs are preferred; when no suitable natural negative exists in moderation data, we alter one precedent-side decision-relevant relation while preserving the query, label, topic, and most lexical content. The legal benchmarks use natural CIGs only. The endpoint targets are USE for $c^{(+)}$ and REJECT for $c^{(-)}$; partial cases supervise LIMIT during structured SFT. Construction details are in Appendix A.8.

Learning objective. Outcome supervision alone does not reveal why a precedent helped or hurt a decision. Two trajectories may obtain the same final label while relying on different precedent relations. We therefore separate task correctness from process-level precedent transfer. The rollout-level task, authority, and relation rewards are defined below.

4.2 GARO AND INTERVENTION-GROUPED OPTIMIZATION

For a precedent-conditioned rollout τ, task reward is

$$R^{\text{task}}(\tau) = \mathbb{I}[\hat{y} = y^\star]. \tag{6}$$

A deterministic checker verifies source consistency, while a frozen label- and verdict-blind critic scores groundedness $G(\tau)$ and materiality $M(\tau)$ on $\{0, 1, 2\}$. Structural authority credit rewards the intended endpoint state (1), the intermediate LIMIT state (0.5), and the opposite endpoint (0). We define

$$R_{\text{rel}}(\tau) = h(\tau)\frac{G(\tau) + M(\tau)}{4}, \qquad R^{\text{proc}}(\tau) = \omega_a R_{\text{auth}}(\tau) + (1-\omega_a)R_{\text{rel}}(\tau), \quad \omega_a = 0.6. \tag{7}$$

Keeping R^{task} and R^{proc} separate avoids *analogical reward aliasing*: two trajectories may receive the same correct outcome while using incompatible precedent relations.

IG-GARO changes the optimization unit from one precedent to the whole intervention group. With $K = 4$ rollouts per condition, task reward is normalized across all $3K$ rollouts and process reward across the $2K$ precedent-conditioned rollouts:

$$A^{\text{IG}}_{q,k} = A^{\text{task}}_{q,k} + \alpha A^{\text{proc}}_{q,k}, \qquad \alpha = 0.5, \qquad A^{\text{proc}}_{0,k} = 0. \tag{8}$$

We optimize this advantage with the standard clipped GRPO objective and a reference KL penalty (full form in Appendix A.7). Single-case GARO uses the same factorized rewards without intervention-group comparison.

4.3 AUTHORITY RANKING AND UTILITY ROUTING

CIGs induce the pairwise authority objective

$$\mathcal{L}_{\text{rank}} = -\mathbb{E}_x \log \sigma\Big(A_\theta(x, c^{(+)}) - A_\theta(x, c^{(-)}) - \gamma\Big), \qquad \gamma = 0.5, \tag{9}$$

and the specialist objective is

$$\mathcal{L}_{\text{specialist}} = \mathcal{L}_{\text{IG-GRPO}} + \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}, \quad \beta = 0.01,\ \lambda_{\text{rank}} = 0.1. \tag{10}$$

For deployment, five-fold cross-fitting provides out-of-fold repair/regression targets $u(x) = r(x) - \lambda_b b(x) - \eta$. The router uses the percentile-calibrated Query-only margin, the top-2 retrieval-score gap, and label entropies among the top-4 and top-8 precedents:

$$v(x, \mathcal{C}_k) = [\tilde{m}_Q(x), \Delta s_{\text{ret}}(x), H_4(x), H_8(x)]. \tag{11}$$

A transparent baseline routes only when the top two historical labels disagree and $\tilde{m}_Q < \tau$. The learned router fits $q_\phi(v)$ to the OOF utility target and invokes the specialist when

$$g_\phi(x, \mathcal{C}_k) = \mathbb{I}[y_{c_1} \neq y_{c_2} \ \wedge \ q_\phi(v(x, \mathcal{C}_k)) > 0]. \tag{12}$$

Cross-fitting, calibration, and runtime details are given in Appendix A.7.

Deployment perspective. Authority governs precedent transfer within the specialist, whereas utility governs whether the specialist is invoked.

5 EXPERIMENTAL SETUP

Datasets. COLD is the primary benchmark (37,480 Chinese comments; 25,726/6,431/5,323 train/dev/test) (Deng et al., 2022). We reserve 800 development queries before model selection for natural authority evaluation. Cross-setting experiments use Ex-ToxiCN-MM (Wang et al., 2026), HED-COLD (Wu et al., 2025), the English USClassActions judgment-prediction benchmark (Semo et al., 2022), and the three-class English CaseFacts precedent benchmark (Putta et al., 2026). Each benchmark uses its own train-only retrieval archive; evaluation identities and near duplicates are excluded before indexing.

Table 1: Main results on the official COLD test set ($N = 5{,}323$). Scores are percentages averaged over three seeds. For always-on precedent-conditioned baselines, Specialist Invocation is 100% by construction; for selective systems, it reports the fraction of queries routed to the specialist branch.

Method	Macro-F1	Accuracy	Conflict F1	Specialist Invocation (%)
Query-only	81.8 ± 0.3	82.1 ± 0.3	72.1 ± 0.5	0.0
Random 4-shot ICL	81.4 ± 0.3	81.7 ± 0.3	73.2 ± 0.5	100
Retrieved-context prompt	81.7 ± 0.3	82.0 ± 0.3	74.9 ± 0.5	100
Flat Case-SFT	81.5 ± 0.3	81.9 ± 0.3	74.4 ± 0.5	100
Outcome-GRPO	82.8 ± 0.4	83.1 ± 0.4	77.0 ± 0.6	100
Class-RAG	82.5 ± 0.3	82.8 ± 0.3	75.1 ± 0.5	100
CARO	83.4 ± 0.3	83.7 ± 0.3	78.0 ± 0.5	100
CHAIRO	84.1 ± 0.3	84.4 ± 0.3	79.3 ± 0.5	100
Single-case GARO	84.0 ± 0.4	84.3 ± 0.4	80.2 ± 0.5	100
Always-on IG-GARO	82.4 ± 0.4	82.7 ± 0.4	85.3 ± 0.5	100
GATE + Single-case GARO	84.7 ± 0.3	85.0 ± 0.3	81.1 ± 0.5	24.8
GATE + IG-GARO	86.5 ± 0.3	86.8 ± 0.3	84.0 ± 0.4	23.2
IG-GARO + UTILITY	$\mathbf{88.1 \pm 0.3}$	$\mathbf{88.5 \pm 0.3}$	$\mathbf{86.2 \pm 0.4}$	**19.7**

Training and models. CIG construction succeeds for 22,400 COLD training queries (87.1% coverage): 10,080 natural matched, 7,840 retrieval-conflict-derived, and 4,480 controlled groups. With three conditions and $K = 4$, one RL pass samples up to 268,800 trajectories. The primary generalist and specialist are Qwen3.5-4B; Gemma-3-4B tests a different backbone and Qwen3.5-14B a larger policy. Qwen3.8-27B is the frozen relation critic and Qwen3.5-35B-A3B generates structured traces and relation signatures. The specialist is initialized from 55,552 filtered structured-SFT traces. Full splits and evaluation protocol are in Appendix A.4; checkpoints, hyperparameters, trace retention, and compute are reported in the remaining appendices.

Audits and metrics. Two 300-example audits evaluate CIG construction and critic groundedness/materiality; the independent authority set contains 800 natural similarity-matched pairs. Macro-F1 is the primary end-to-end metric. Conflict F1 is reported on top-2 label-disagreement queries. Specialist invocation rate is 100% by construction for always-on precedent-conditioned methods; for selective systems, it is the fraction of queries routed to the specialist branch. Authority discrimination accuracy (ADA) measures whether $A_\theta(x, c^{(+)}) > A_\theta(x, c^{(-)})$; strict transfer-state accuracy additionally requires USE/REJECT. Primary results use three seeds and paired bootstrap confidence intervals.

6 RESULTS

Research questions. We organize experiments around four questions. **RQ1** evaluates end-to-end performance improvements. **RQ2** examines whether IG-GARO captures authority rather than retrieval or label shortcuts. **RQ3** analyzes when precedent reasoning is useful and how routing controls its cost. **RQ4** studies transfer across domains and model settings.

To examine whether authority learning depends on controlled interventions, we additionally evaluate natural CIGs constructed from retrieval conflicts without relation editing. Natural-only CIGs achieve 85.9 ADA, recovering most of the full intervention-based result (87.6 ADA), suggesting that IG-GARO learns transferable applicability boundaries rather than relying on intervention-specific artifacts.

6.1 RQ1: OVERALL EFFECTIVENESS

Prompting and flat Case-SFT remain near Query-only, while outcome-only GRPO gives a modest gain. Always-on IG-GARO improves Conflict F1 to 85.3 but yields only a small overall gain (82.4 vs. 81.8 Macro-F1), indicating that its benefit depends on the query condition. Gate + IG-GARO reaches 86.5 Macro-F1 versus 84.7 for Gate + Single-case GARO, and IG-GARO + Utility further

raises the score to 88.1 while invoking the specialist on 19.7% of queries. The training-stage and routing-stage improvements are both significant (95% CIs $[+1.1, +2.5]$ and $[+1.0, +2.2]$).

6.2 RQ2: Learning Decision Authority

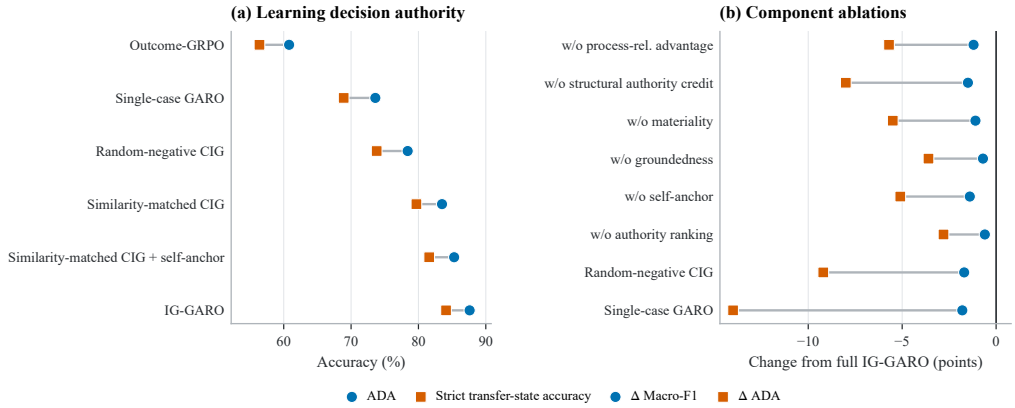


Figure 3: Learning decision authority and component ablations. (a) Authority discrimination accuracy (ADA) and strict transfer-state accuracy across training designs. (b) Changes in Macro-F1 and ADA relative to full IG-GARO under the same deployment rule; more negative values indicate larger performance drops when a component is removed or replaced.

CIG training produces a much larger change in authority discrimination than in ordinary label accuracy. ADA rises from 73.6 with Single-case GARO to 87.6 with IG-GARO. Similarity-matched CIGs are more informative than random negatives (83.5 vs. 78.4 ADA), and IG-GARO retains 84.7 ADA even when the positive and negative precedents share the same historical label. Natural-only CIGs reach 85.9 ADA, close to the 87.6 obtained with the full construction.

The human audits are consistent with these results: constructed CIG targets agree with adjudication in 91.3% of cases, while critic–human agreement is higher for groundedness ($\kappa = 0.80$) than for materiality ($\kappa = 0.72$). Allowing LIMIT cases to carry authority lowers Macro-F1 from 88.1 to 87.3, indicating that partial applicability is not sufficient for decision transfer. Appendix Fig. 6 gives a candidate-level view, and relation-level and sensitivity analyses are reported in Appendix A.1.

6.3 RQ3: Conditional Expertise and Selective Deployment

The value of the specialist is highly conditional. Under top-2 retrieval disagreement, always-on IG-GARO improves Accuracy by 13.5 points and Macro-F1 by 13.2 points; under agreement, the corresponding changes are -4.6 and -4.7. Since agreement queries make up most of the test set, these effects largely cancel under always-on deployment. Routing turns this conditional advantage into an end-to-end gain. Gate + IG-GARO reaches 86.5 Macro-F1 with a 23.2% call rate, and IG-GARO + Utility reaches 88.1 with 19.7%. The latter reduces mean/P95 latency from 5.95/8.62 s under always-on IG-GARO to 2.72/5.98 s (Appendix A.12). OOF, final-development, and test call rates differ by less than 0.7 points; additional routing controls are reported in Appendix A.2.

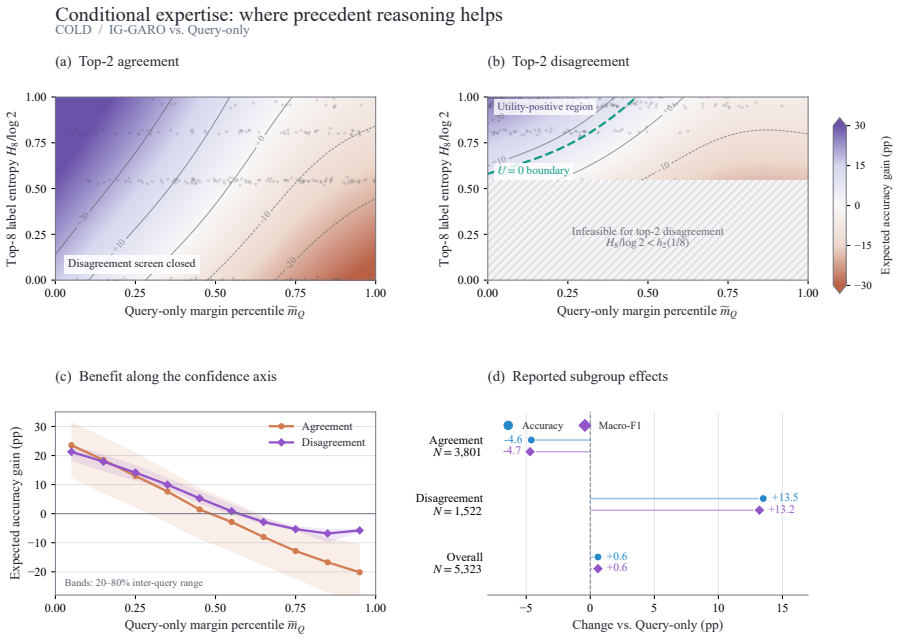


Figure 4: Conditional value of precedent reasoning on the official COLD test set. (a–b) Specialist-minus-Query-only expected accuracy gain over calibrated Query-only margin and top-8 precedent-label entropy, separated by top-2 label agreement and disagreement; the hatched region in (b) is infeasible under top-2 disagreement. (c) Conditional gain along the Query-only confidence axis, with bands showing the 20–80% inter-query range. (d) Observed subgroup changes in Accuracy and Macro-F1; positive values favor always-on IG-GARO over Query-only.

6.4 RQ4: Robustness Across Settings

Table 2: Cross-setting and cross-domain Macro-F1 across moderation and legal reasoning tasks. CARO and CHAIRO are included as analogy-based case reasoning baselines. Each row trains on the corresponding benchmark's training partition.

Dataset / backbone	Query-only	CARO	CHAIRO	Single-case GARO	Gate + IG-GARO	IG-GARO + Utility
COLD / Qwen3.5-4B	81.8	83.4	84.1	84.0	86.5	**88.1**
Ex-ToxiCN-MM / Qwen3.5-4B	78.9	85.6	86.2	81.5	84.8	**86.7**
HED-COLD / Qwen3.5-4B	76.4	83.7	83.3	79.6	82.7	**84.6**
COLD / Gemma-3-4B	78.3	84.5	84.2	80.7	83.6	**85.2**
COLD / Qwen3.5-14B	85.6	87.7	87.4	87.2	88.8	**89.9**
USClassActions (EN, legal) / Qwen3.5-4B	61.9	68.0	68.6	64.8	67.2	**69.1**
CaseFacts (EN, legal, 3-class) / Qwen3.5-4B	56.4	62.6	62.9	59.6	62.7	**64.9**

Across all seven settings in Table 2, IG-GARO + Utility has the highest Macro-F1 among the compared methods. The gains are not confined to COLD or to the Qwen3.5-4B backbone: the same pattern appears on additional moderation benchmarks, a different backbone, a larger model, and two independently trained English legal tasks. These experiments measure robustness across settings rather than zero-shot transfer. Representative training trajectories are shown in Appendix Fig. 7.

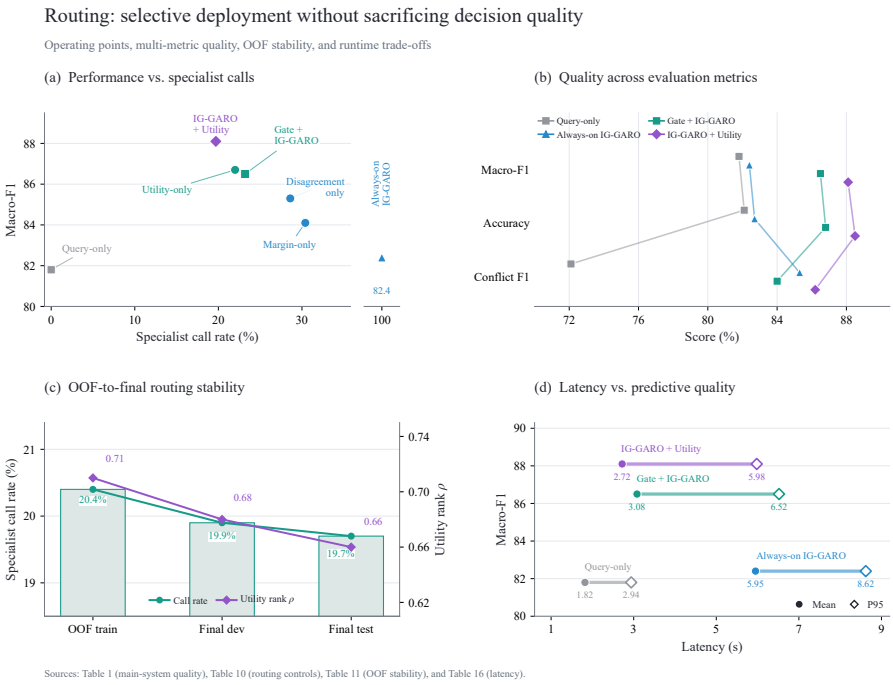


Figure 5: Deployment diagnostics with a fixed IG-GARO specialist. (a) Macro-F1 versus specialist invocation rate across routing controls. (b) Macro-F1, Accuracy, and Conflict F1 for the main system variants. (c) OOF-to-final stability of specialist invocation rate and utility-rank correlation. (d) Predictive quality versus end-to-end latency; markers denote mean and P95 latency.

7 Discussion

The experiments point to a distinction that retrieval similarity alone does not capture. Cases with similar wording can differ sharply in whether their historical decisions are transferable, and similarity-matched CIGs provide a direct training signal for that boundary. The same effect remains visible when positive and negative precedents share a label, making simple label copying an unlikely explanation for the authority gains.

Precedent reasoning is also not uniformly beneficial. IG-GARO is most useful when retrieved cases disagree and the Query-only prediction is uncertain; on many easier queries, invoking the specialist can instead introduce errors. This explains why the always-on specialist improves conflict cases without producing a comparable overall gain. Authority learning and utility routing address different parts of this problem: the former controls which retrieved decisions can transfer, and the latter controls when precedent-conditioned reasoning is worth using.

8 Conclusion

Case retrieval raises a question that relevance alone cannot answer: when should a retrieved decision carry over to a new query? We formulate this as decision-authority learning and use similarity-matched contrastive interventions to expose the transfer boundary. IG-GARO learns this boundary by separating task correctness from precedent-transfer quality, while utility routing limits specialist use to queries where it is expected to help. Across moderation and legal reasoning settings, the results favor treating precedent transfer as a learned decision rather than a consequence of semantic similarity.

Ethics Statement

The work uses existing moderation and legal datasets under their distribution terms. Controlled perturbations are training interventions rather than natural observations, and potentially harmful content requires appropriate annotation and review.

Reproducibility Statement

The appendix reports schemas, CIG provenance, cross-fitting, optimization details, audit rubrics, compute, and statistical procedures; retrieval archives and controlled interventions use training partitions only.

AI Assistance Disclosure

Generative AI tools were used to assist with the writing and presentation of this manuscript, including language polishing, improving clarity and concision. All AI-assisted edits were reviewed by the authors for technical accuracy, consistency with the reported experiments, and faithfulness to the intended scientific claims. The authors take full responsibility for the final content of the paper.

References

Kevin D. Ashley. *Modeling Legal Arguments: Reasoning with Cases and Hypotheticals*. MIT Press, 1991.

Jianfa Chen, Emily Shen, Trupti Bavalatti, Xiaowen Lin, Yongkai Wang, Shuming Hu, Harihar Subramanyam, Ksheeraj Sai Vepuri, Ming Jiang, Ji Qi, Li Chen, Nan Jiang, and Ankit Jain. Class-RAG: Real-time content moderation with retrieval augmented generation. *arXiv preprint arXiv:2410.14881*, 2024.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11580–11599. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.796. URL https://aclanthology.org/2022.emnlp-main.796.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out*, pp. 100–114, 2022. doi: 10.18653/v1/2022.deelio-1.10.

Haotian Lu, Yuchen Mou, and Bingzhe Wu. CHAIRO: Contextual hierarchical analogical induction and reasoning optimization for LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 36526–36537, 2026. doi: 10.18653/v1/2026.acl-long.1692.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759.

Akshith Reddy Putta, Jacob Devasier, and Chengkai Li. CaseFacts: A benchmark for legal fact-checking and precedent retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17246–17265. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.785. URL https://aclanthology.org/2026.acl-long.785/.

Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial Intelligence*, 49(1–3):361–395, 1991.

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. ClassActionPrediction: A challenging benchmark for legal judgment prediction of class action cases in the us. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pp. 31–46. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.nllp-1.3. URL https://aclanthology.org/2022.nllp-1.3/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Weiming Wang, Junyu Lu, Han Wang, Xiaokun Zhang, Zewen Bai, Bo Xu, Liang Yang, and Hongfei Lin. Distinguishing right from wrong in debates: Attribution analysis of chinese harmful memes. *arXiv preprint arXiv:2605.24344*, 2026.

Bingzhe Wu, Haotian Lu, and Yuchen Mou. CARO: Chain-of-analogy reasoning optimization for robust content moderation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28880–28892, 2026. doi: 10.18653/v1/2026.findings-acl.1442.

Junqi Wu, Shujie Ji, Kang Zhong, Huiling Peng, Zhendong Xiao, Xiongding Liu, and Wu Wei. Enhancing Chinese offensive language detection with homophonic perturbation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22660–22675. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1154. URL https://aclanthology.org/2025.emnlp-main.1154/.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2422–2437, 2022. doi: 10.18653/v1/2022.emnlp-main.155.

A Appendix

A.1 Additional Authority Analyses

The following figure and tables provide a candidate-level view of authority, the natural-CIG comparison, relation-type breakdown, hyperparameter sensitivity, and human audits summarized in RQ2.

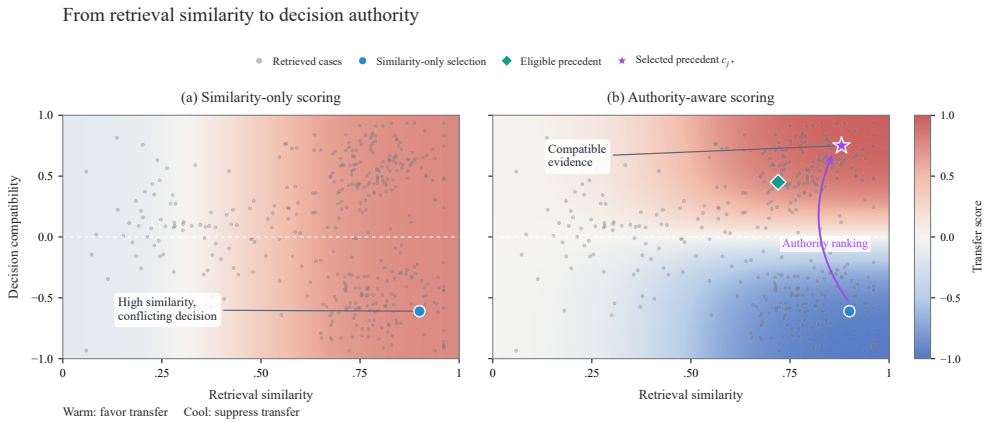


Figure 6: Retrieval similarity and learned decision authority for candidate precedents. Similarity-only selection can favor a highly similar candidate whose historical decision should not transfer. Under the learned specialist, D_θ determines eligibility and A_θ ranks eligible precedents; $c_{j^\star}$ is the selected precedent in Eq. (3).

Table 3: Composition of contrastive intervention groups used for authority learning. Natural conflicts provide observed applicability boundaries, while controlled interventions increase coverage of decision-relevant relations.

CIG source	Number of groups	Ratio (%)
Natural matched conflicts	10,080	45.0
Retrieval-conflict derived groups	7,840	35.0
Controlled relation interventions	4,480	20.0
Total	22,400	100.0

Table 4: Authority learning with increasingly informative specialist training designs. Macro-F1 is reported under transparent-gate deployment for each specialist; ADA evaluates the underlying specialist on the natural authority set.

Specialist training design	Gated Macro-F1	ADA
Single-case GARO	84.7	73.6
Random-negative CIG	84.8	78.4
Natural-only matched CIG	86.0	85.9
Full CIG (natural + controlled)	**86.5**	**87.6**

Table 5: Authority discrimination by material-relation type on the 800-pair natural authority set.

Relation type	N	Single-case GARO	IG-GARO
Stance	168	75.6	**89.3**
Quotation	121	77.7	**90.1**
Endorsement / rejection	174	74.1	**88.5**
Target	137	72.3	**86.9**
Speech act	96	71.9	**85.4**
Contextual role	104	68.3	**83.7**
Overall	800	73.6	**87.6**

Table 6: Sensitivity of the core authority-learning controls on COLD. Macro-F1 is measured under the fixed transparent gate; ADA uses the 800-pair authority set.

Parameter	Setting	Macro-F1	ADA	Relative rollout cost
α	0.00	85.3	81.9	1.00×
	0.25	86.1	85.4	1.00×
	0.50	**86.5**	**87.6**	1.00×
	0.75	86.3	87.2	1.00×
	1.00	85.8	86.4	1.00×
ω_a	0.20	85.8	84.4	1.00×
	0.40	86.2	86.5	1.00×
	0.60	**86.5**	**87.6**	1.00×
	0.80	86.2	86.8	1.00×
	1.00	85.6	83.9	1.00×
K	2	86.0	85.8	0.51×
	4	**86.5**	**87.6**	1.00×
	8	86.6	87.9	1.98×

Table 7: Human audit of structural labels on 300 sampled constructed CIGs.

Audit measure	Value
Human–human weighted κ	0.82 $[0.77, 0.87]$
Constructed target vs. adjudication	91.3% $[87.6, 94.0]$
$c^{(+)}$ applicability validity	92.0% $[88.4, 94.6]$
$c^{(-)}$ non-applicability validity	90.7% $[86.9, 93.5]$

Table 8: Human audit of the frozen relation critic on 300 held-out traces.

Signal	Human–human κ	Critic–human κ	Macro-F1
Groundedness	0.84 $[0.79, 0.88]$	0.80 $[0.75, 0.84]$	89.1 $[85.1, 92.1]$
Materiality	0.78 $[0.72, 0.83]$	0.72 $[0.65, 0.78]$	83.4 $[78.8, 87.2]$

A.2 Additional Routing Analyses

The following tables provide the conditional subgroup breakdown, routing operating points, and OOF-to-final call-rate stability summarized in RQ3. All routing variants use the same fixed IG-GARO specialist; the reported invocation rate is the fraction of test queries routed to that specialist.

Table 9: Conditional specialist behavior by retrieval state. Values are percentages; Δ compares always-on IG-GARO with Query-only within the same subset.

	Accuracy			Macro-F1		
State	QO	IG	Δ	QO	IG	Δ
Top-2 agreement ($N = 3{,}801$)	85.9	81.3	-4.6	85.6	80.9	-4.7
Top-2 disagreement ($N = 1{,}522$)	72.7	86.2	$+13.5$	72.1	85.3	$+13.2$
Overall ($N = 5{,}323$)	82.1	82.7	$+0.6$	81.8	82.4	$+0.6$

Table 10: Routing operating points with a fixed IG-GARO specialist. Invocation is the percentage of test queries routed to the specialist.

Routing rule	Macro-F1	Specialist Invocation (%)
Always-on specialist	82.4	100.0
Margin-only	84.1	30.4
Disagreement-only	85.3	28.6
Transparent Gate	86.5	23.2
Utility-only	86.7	22.0
Disagreement + Utility	**88.1**	**19.7**

Table 11: Routing stability from out-of-fold supervision to final checkpoints.

Evaluation source	Call rate (%)	Utility rank ρ
OOF training predictions	20.4	0.71
Final development	19.9	0.68
Final test	19.7	0.66

A.3 Additional Cross-setting Results

Figure 7 complements the endpoint comparison in Table 2 with training trajectories for four representative dataset/backbone settings. The trajectories are diagnostic: the main cross-setting claim is based on the final Macro-F1 values reported in Table 2.

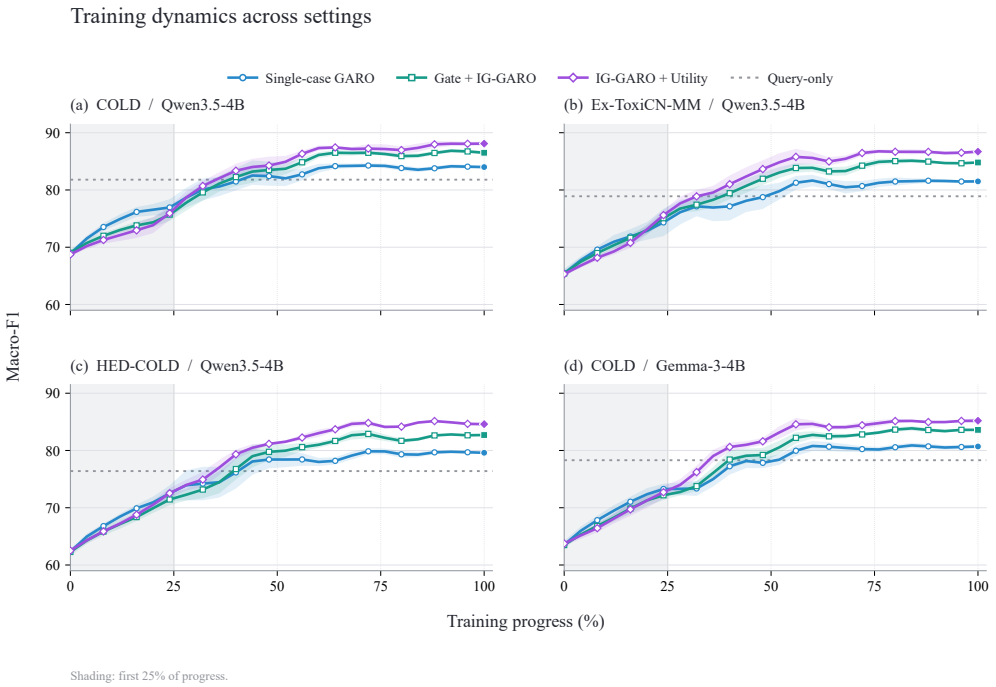


Figure 7: Training dynamics across representative datasets and backbones. Curves report Macro-F1 over normalized training progress for Single-case GARO, Gate + IG-GARO, and IG-GARO + Utility; dashed horizontal lines denote Query-only performance. Final values correspond to the settings reported in Table 2. Five recorded runs are used to visualize training dynamics; statistical endpoint comparisons use the three prespecified complete method seeds described in Section 5.

A.4 Dataset and Evaluation Details

COLD uses the official 25,726/6,431/5,323 train/dev/test split; 800 development queries are reserved before model selection for authority evaluation, leaving 5,631 for hyperparameter selection. ToxiCN-MM uses 4,923/696/1,423, and HED-COLD uses 6,300/700/3,000. USClassActions uses a stratified 2,100/300/600 split. CaseFacts uses 5,214/580/500, with the development holdout grouped by `fact_id`. Each benchmark has its own train-only retrieval archive; development/test identities, exact duplicates, and near duplicates are excluded before indexing, and secondary benchmarks never reuse COLD precedents.

The 300-group CIG audit is independently labeled for positive-case full applicability, negative-case non-applicability, and material relation type, with adjudication by a third annotator. A separate 300-trace audit evaluates critic groundedness and materiality. The 800-pair authority set uses natural similarity-matched precedents from the training archive and contains no controlled interventions.

A.5 Additional Cheap Case-Augmentation Control

Table 12: Additional low-cost controls on COLD.

Method	Macro-F1	Accuracy	Conflict F1
Top-k label vote	75.2	75.9	66.1
QO + top-2 labels only	80.9	81.3	72.6

Simple label voting and label-only prompting underperform both Query-only and full retrieved-context prompting.

A.6 Structured Output and Verification

The precedent-conditioned specialist returns:

```
{
  "precedent_id": "...",
  "query_quote": {"field": "...", "quote": "..."},
  "case_quote": {"case_id": "...", "field": "...", "quote": "..."},
  "similarity": "...",
  "material_difference": "...",
  "transfer": "USE|LIMIT|REJECT",
  "reason": "...",
  "label": "..."
}
```

`precedent_id` identifies the precedent being evaluated by the candidate-conditioned interface. The transfer field records relation applicability: USE grants decision authority, LIMIT records partial applicability without decision transfer, and REJECT excludes the precedent. The no-precedent self-anchor returns only a reason and label. Case identity, nonempty exact spans, allowed source fields, transfer states, and task-label parsing are deterministic checks. The critic receives evidence and relation/transfer fields only after deterministic checking; gold query labels, final student verdicts, and reward values are removed.

A.7 Implementation Details

IG-GRPO objective. For each CIG, we sample $\mathcal{G}_x = \{\tau_k^{(q)} : q \in \{0, +, -\}, k = 1, \ldots, K\}$. Task advantages are normalized over all $3K$ rollouts, while process advantages are normalized over the $2K$ precedent-conditioned rollouts:

$$A_{q,k}^{\text{task}} = \frac{R^{\text{task}}(\tau_k^{(q)}) - \mu_T}{\sigma_T + \epsilon_A}, \qquad A_{q,k}^{\text{proc}} = \frac{R^{\text{proc}}(\tau_k^{(q)}) - \mu_P}{\sigma_P + \epsilon_A}.$$

For the self-anchor, $A^{\text{proc}}_{0,k} = 0$. If a component has zero variance, its normalized advantage is set to zero. For response token t, with current-to-rollout policy ratio $\rho_{q,k,t}(\theta)$,

$$\mathcal{L}_{\text{IG-GRPO}} = -\mathbb{E}_{x,q,k}\left[\frac{1}{T_{q,k}}\sum_{t=1}^{T_{q,k}} \min\big(\rho_{q,k,t}A^{\text{IG}}_{q,k}, \text{clip}(\rho_{q,k,t}, 1-\epsilon, 1+\epsilon)A^{\text{IG}}_{q,k}\big)\right].$$

Router calibration and features. For OOF predictions,

$$r(x) = \mathbb{I}[f^{\text{OOF}}_Q(x) \neq y^\star, f^{\text{OOF}}_R(x) = y^\star], \quad b(x) = \mathbb{I}[f^{\text{OOF}}_Q(x) = y^\star, f^{\text{OOF}}_R(x) \neq y^\star].$$

We percentile-calibrate the Query-only top-two logit margin $\tilde{m}_Q = F_Q(m_Q)$. If $p_k(y)$ is the empirical label frequency among the top-k precedents, $H_k = -\sum_y p_k(y)\log p_k(y)$. The utility MLP is trained with SmoothL1 loss on $u(x) = r(x) - \lambda_b b(x) - \eta$. The transparent gate uses $d(x) = \mathbb{I}[y_{c_1} \neq y_{c_2}]$ and routes when $d(x) = 1$ and $\tilde{m}_Q < \tau$.

Table 13: Default configuration for decision-authority learning.

Component	Configuration
Primary backbone	Qwen3.5-4B
Cross-backbone / scale	Gemma-3-4B / Qwen3.5-14B
Frozen relation critic	Qwen3.8-27B, deterministic decoding
Interventions/query	self-anchor, full-applicability endpoint, non-applicability endpoint
Rollouts/condition	$K = 4$; 12/CIG
Process composition	$\omega_a = 0.6$ (authority), 0.4 (relation)
Factorized advantage	task over $3K$; process over $2K$; $\alpha = 0.5$
Policy clipping / KL	$\epsilon = 0.2, \beta = 0.01$
Authority ranking	$\gamma = 0.5, \lambda_{\text{rank}} = 0.1$
Transparent gate	$d(x) = 1$ and $\tilde{m}_Q(x) < \tau$; τ fixed on development
Utility estimator	$4 \to 32 \to 16 \to 1$ MLP
Utility penalties	$\lambda_b = 1.5, \eta = 0.05$
Utility supervision	five-fold OOF branch outcomes
Query-only margin	empirical-CDF percentile

Cross-fitting boundary. For each of five folds, temporary branch checkpoints, retrieval archives, teacher traces, and CIGs are constructed from the corresponding in-fold training partitions only. Held-out queries supply branch outcomes and features, not training precedents. Final generalist and IG-GARO checkpoints are fit on the complete training partition after OOF supervision is fixed.

Runtime procedure. Run the generalist and retrieve identity-distinct precedents. The transparent Gate baseline calls IG-GARO when the top two historical labels disagree and the calibrated Query-only margin falls below τ. The learned router uses the same disagreement screen and calls IG-GARO when $q_\phi(v) > 0$. Once called, the specialist evaluates each candidate's transfer state, keeps only USE candidates in the authority set, and selects the highest-scoring one. If the set is empty, it returns the self-anchor prediction. Invalid specialist output falls back to the generalist and still counts as a call. The relation critic is not used online.

A.8 CIG PROVENANCE AND AUDIT

The construction teacher emits task-specific relation signatures. Moderation uses target, stance, endorsement/rejection, quotation status, speech act, and contextual role. Legal judgment prediction uses cause of action, alleged conduct, procedural posture, jurisdiction, and requested remedy. Legal fact checking uses legal issue, holding polarity, temporal validity, support/contradiction relation, and authority status. These signatures are construction signals, not gold evaluation labels.

Each CIG record stores query identity, positive/negative precedent identities, retrieval scores, relation signatures, differing material relation, structural target, and construction provenance. Before indexing, normalized exact-text hashes remove duplicates; a candidate pair is treated as near-duplicate when BGE-M3 cosine similarity exceeds 0.985 and character 5-gram Jaccard similarity exceeds

0.90. Source-identical pairs are excluded independently of these thresholds. Natural matched pairs use a development-fixed retrieval-score tolerance of 0.03. Retrieval-conflict-derived groups are natural candidate sets whose top retrieved historical labels disagree before relation filtering. Natural matched pairs, retrieval-conflict-derived pairs, and controlled relation interventions use distinct tags. Controlled edits preserve topic and most lexical content but are not represented as natural observations.

The authority set uses the 800 development queries reserved before model selection. Each query is paired with natural similarity-matched candidates from the training archive; no controlled interventions are used. Two annotators independently assign applicability and material-relation labels before adjudication. The separate CIG audit samples 300 constructed training groups to estimate construction-label noise.

Cross-domain CIGs. USClassActions and CaseFacts use natural matched CIGs only. The legal relation schemas are instantiated from the task fields described in Section 4.1, and candidate precedents are retrieved exclusively from each benchmark's training archive. For CaseFacts, the original `fact_id` grouping is preserved before the development holdout so that claim/negation pairs cannot cross archive boundaries.

A.9 Critic Audit Rubric

Groundedness score 0 denotes unsupported or source-misattributed relations; score 1 denotes partly supported or insufficiently specific relations; score 2 denotes source-supported similarities and differences. Materiality score 0 denotes superficial overlap or irrelevant distinctions; score 1 denotes a plausible but incomplete applicability distinction; score 2 denotes a clear decision-relevant distinction concerning target, endorsement, quotation, speech act, stance, or contextual role.

Two independent annotators score all 300 sampled traces before adjudication. A third annotator resolves disagreements without replacing the two original ratings used for human–human agreement. Weighted ordinal κ and macro-F1 are reported with bootstrap confidence intervals.

A.10 Structured-SFT Trace Retention

Table 14: Structured-SFT teacher-trace filtering statistics.

Partition	Field	Value
Overall	Generated traces	64,000
Overall	Passed deterministic filtering	55,552
Overall	Retention	86.8%
Task label	Offensive	86.4%
Task label	Safe	87.2%
Transfer state	Full applicability / USE	87.1%
Transfer state	Non-applicability / REJECT	86.5%
Transfer state	Partial applicability / LIMIT	85.9%

These traces form the fixed structured-SFT initialization pool and are not counted as IG-GARO rollout trajectories. The latter are sampled on-policy from CIGs during reinforcement learning, with up to 268,800 trajectories per complete pass under the default configuration. The most common SFT-trace rejection reasons are invalid source spans, case-ID mismatch, malformed schema, action parse failure, and duplicate-source conflicts.

A.11 TRAINING COST

Table 15: Representative training cost for the COLD/Qwen3.5-4B pipeline on H100-80GB GPUs. GPU-hours aggregate device count times wall-clock duration.

Stage	GPUs	Wall time (h)	GPU-hours
Structured SFT	8	3.2	25.6
Single-case GARO	16	10.1	161.6
IG-GARO	32	11.9	380.8
Five-fold OOF branch generation	16	13.4	214.4
Utility MLP fitting	1	0.2	0.2
Full decision-authority pipeline	–	–	621.0

The 27B relation critic and 35B-A3B teacher are offline-only. Online inference retains the 4B generalist/specialist stack, so the added training cost is exchanged for selective rather than always-on specialist use at deployment.

A.12 EFFICIENCY

Table 16: End-to-end inference efficiency on H100-80GB. Specialist invocation rate is reported as the percentage of queries that invoke the specialist.

System	Mean latency (ms)	P95 (ms)	Specialist Invocation (%)
Query-only	1,820	2,940	0
Always-on IG-GARO	5,950	8,620	100
Gate + IG-GARO	3,080	6,520	23.2
IG-GARO + Utility	**2,720**	**5,980**	**19.7**

A.13 STATISTICAL REPORTING

All primary rows use three complete method seeds on the same frozen test identities. Per-seed task metrics and specialist invocation rates are computed before reporting means and sample standard deviations.