Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.30609

Large Model Semantic Parsing of Court Records with Bias Calibration Benchmark for Restorative Justice

Cunhui Li

Faculty of Law, The University of New South Wales, Sydney, Australia
licunhui02371@outlook.com

Abstract. Restorative justice requires algorithms that are simultaneously accurate, explainable, and fair in reconstructing facts and repairing relationships, yet domestic court records exhibit cross-domain heterogeneity, weakly explicit procedural cues, and proxy signals for sensitive attributes, leaving current legal NLP with gaps in event-chain integrity, structural verifiability, and subgroup bias control. To address these challenges, an ontology-driven approach centered on constrained structural decoding and explicit structural validation under an "events–roles–measures–outcomes" schema is proposed, augmented with statute/analogous-case retrieval to stabilize cross-domain parsing, and coupled with an RJ-BCB evaluation–calibration loop comprising counterfactual generation, cost-sensitive reweighting, and groupwise/temperature recalibration, executed and reproduced on domestic public and compliant corpora from 2019–2024. Results show consistent improvements over strong baselines: Slot-F1 rises to 88.9 (+4.7), Span-F1 to 86.3 (+4.6), Graph-F1 to 82.4 (+5.9), and Structural Validity Rate (SVR) to 96.7 (+7.6); under attribute-visible/attribute-hidden protocols, Conditional Opportunity Gap (COG) drops to 2.1%/1.8%, Equalized Odds Difference (EOD) to 2.9%/2.5%, and Counterfactual Invariance Score (CIS) increases to 0.91/0.90, with intergroup calibration error markedly reduced. The significance lies in the synergy between structural constraints and retrieval evidence, which jointly enhances parsing coverage and structural correctness, while the calibration loop effectively compresses disparities driven by non-causal sensitive attributes. The framework provides a measurable, calibratable, and verifiable technical pathway for evidence-chain presentation and recommendation review in restorative-justice settings, and establishes methodological and tooling foundations for compliant deployment and continuous auditing of judicial AI.

Keywords: Semantic Parsing, Legal NLP, Large Language Models, Bias Calibration, Restorative Justice

1. Introduction

Restorative justice seeks to rebuild relationships among victims, offenders, and communities through dialogue, responsibility, and reparative measures, which imposes simultaneous requirements of accuracy, explainability, and fairness on judicial AI. As large models permeate legal NLP, semantic parsing of court records is shifting from flat element extraction to ontology- and event-framed structured generation [1]. However, cross-domain heterogeneity, sparse supervision, and

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.30609

weakly explicit procedural cues impair robustness and verifiability, while leakage of sensitive attributes accumulates group disparities along the "parse–retrieve–recommend" pipeline [2]. Despite advances in legal pretraining, statute matching, and argument extraction, task-coupled bias metrics, counterfactual stability tests, and subgroup calibration strategies tailored to restorative contexts remain underdeveloped. This paper introduces an ontology-grounded structured parsing framework combined with counterfactual augmentation, cost-sensitive reweighting, and post-hoc recalibration to form a closed-loop intervention, and releases a restorative justice benchmark spanning parsing quality, evidence-chain consistency, and bias indicators to enable a measurable, calibratable, and verifiable parsing pathway.

2. Literature review

2.1. Restorative justice and court records

Restorative justice prioritizes procedural fairness and relational repair, which requires contextualized understanding of factual elements, procedural stages, and outcomes; court records offer latent structure yet exhibit heterogeneous narrative practices, uneven argumentative granularity, and jurisdictional variations that induce multi-modal realizations of the same legal constructs [3]. Prior efforts rely on element lists and template extraction to support aggregate comparisons, but under-represent the coupling among procedural sentences, balancing reasons, and unstructured corroboration, limiting evidence-chain display suited to reparative aims [4]. Moreover, sensitive attributes often surface indirectly; without controlled views and counterfactual pairing, parsers conflate correlation with causation and shift recommendation intensity systematically, generating tension, rather than convergence, with restorative objectives.

2.2. Legal NLP and semantic parsing

Legal NLP has progressed from rule-based and bag-of-words systems to domain-pretrained, multitask architectures, improving element recognition across statute retrieval, argument extraction, and fact-to-article mapping, yet cross-domain transfer remains vulnerable to label sparsity and expression drift, yielding structural mismatches [5]. Recent work employs event ontologies, hierarchical labels, and constrained decoding to ensure verifiable outputs, while retrieval augmentation injects statutes and analogous cases to stabilize reasoning consistency; however, without end-to-end constraints that bind procedural pragmatics and evidence associations, models still suffer from slot omissions, relation ambiguity, and inferential drift [6]. For restorative objectives, parsers must elevate measure types, victim participation, and repair outcomes to first-class targets and feed structured outputs directly into evidence-chain visualization and recommendation modules to avoid error amplification and blurred accountability from secondary mappings.

2.3. Bias, calibration, and evaluation in LLMs

Social biases learned by general LLMs surface in legal settings as asymmetric error rates and confidence misalignment across groups, especially when sensitive attributes appear implicitly via semantic adjacency or contextual proxies, where average metrics mask tail-risk harms [7]. Existing bias evaluations emphasize generic toxicity or stereotypes and misalign with the causal structure of legal tasks; likewise, plain temperature scaling improves global calibration but may not reduce subgroup conditional opportunity gaps. An effective route couples counterfactual generation and

DOI: 10.54254/2755-2721/2026.30609

paired testing that flips controlled attributes without perturbing legally relevant causal paths, while deploying cost-sensitive reweighting, groupwise recalibration, and invariance regularization across training and inference [8]. Evaluation should jointly track evidence-chain consistency and counterfactual stability so that the "measure–calibrate–verify" loop remains structurally aligned with restorative objectives.

3. Experimental methods

3.1. Data collection

The corpus is confined to domestic public sources and compliant partnership channels, covering the Supreme People's Court's open judgments archive (and mirrored repositories), provincial high courts, and municipal/intermediate and basic courts' public collections, together with prosecutorial sentencing recommendations and courtroom opinions from People's Procuratorates, and civil mediation agreements subject to judicial confirmation. The temporal window spans 2019–2024, with a pilot extension in the first half of 2025 for out-of-sample validation. Case categories focus on four domains closely tied to reparative dispositions: juvenile matters, domestic violence, property offenses, and minor violent offenses [9]. Documents preserve core segments, factual narratives, procedural stages, statutory citations, reasoning sections, and dispositions, and undergo uniform page-layout parsing and paragraph segmentation.

Processing follows an auditable pipeline. De-identification applies consistent replacements for names, addresses, contact details, and employing entities, and constructs controlled views of sensitive attributes including gender, age, ethnicity, and socioeconomic status. An expert-agreed hierarchical ontology and slot schema, events, roles, measures, and outcomes, are then built, with simultaneous annotation of evidence-chain pointers and temporal order; only items achieving high inter-annotator agreement via dual annotation with arbitration are admitted to the training pool. To support bias analysis, structure-preserving counterfactual instances are generated by flipping non-causal sensitive attributes while holding facts and statutes fixed. A stratified split by court and year produces training, development, and test subsets. Both JSONL and graph representations are released to enable direct structural validation and traceability of downstream parsing outputs.

3.2. Model and training objectives

The model is an instruction-aligned large language model equipped with ontology-constrained, grammar-based decoding and a structural validator. Inputs are de-identified, segmented record passages; outputs are JSON and graph structures conforming to the ontology, with on-the-fly structural validity checks that reject illegal fields and relations [10]. Let (x,y) denote a sample and (x^{cf},y) its counterfactual construction; let w_c be class weights. As shown in Equation (1).

$$\mathcal{L}=\sum_t \mathrm{CE}(p_\theta(y_t|x),y_t)w_{c(t)}+\lambda_{inv}\|f_\theta(x)-f_\theta(x^{cf})\|_2^2+\lambda_{str}1_{[-\mathrm{valid}(o_\theta)]} \quad (1)$$

Here p_θ is the conditional distribution, CE is cross-entropy, and y_t is the label of the t-th slot or relation. The term $w_{c(t)}$ mitigates class imbalance. The function $f_\theta(\cdot)$ encodes a joint representation of measure intensity and key slots. The decoded output is o_θ, and valid$(\cdot)$ denotes structural compliance with the ontology. Scalars λ_{inv} and λ_{str} weight the counterfactual invariance penalty and structural penalty respectively.

DOI: 10.54254/2755-2721/2026.30609

Confidence calibration uses post-hoc temperature scaling on the development set. Let z_{ik} be the logit of sample i for class k; the optimal temperature satisfies [11]. As shown in Equation (2).

$$T^{*}=\arg\min_{T>0}\sum_{i}-\log\frac{\exp(z_{i,y_i}/T)}{\sum_{k}\exp(z_{ik}/T)} \tag{2}$$

The global temperature T reduces overconfidence and improves subgroup-level confidence–frequency alignment. Statutory provisions and analogous-case evidence are retrieved and injected as visible context to stabilize structural consistency across regions and years, ensuring that decoding remains verifiable under distributional shifts.

3.3. Benchmark and metrics

The RJ-BCB benchmark comprises three task families with reproducible measurement procedures. Parsing quality includes four metrics: Slot-F1 for macro-averaged slot recognition; Span-F1 for boundary and label matching of key spans; Graph-F1 counting a hit only when nodes and relations are simultaneously correct; and Structural Validity Rate (SVR), the proportion of outputs that satisfy ontology and decoding constraints. Evidence-chain consistency includes two metrics: Event-chain Consistency Rate (ECR), which assesses order-preserving alignment of temporal and causal dependencies; and Rationale–Conclusion Alignment (RCA), which verifies whether cited statutes and reasoning segments substantiate the recommended measures and outcomes, aggregated at the case level.

Bias calibration covers four metrics: Conditional Opportunity Gap (COG) compares hit-rate disparities between sensitive subgroups under the same ground-truth labels; Equalized Odds Gap (EOD) jointly measures subgroup differences in false-positive and false-negative rates; Groupwise Calibration Error (GCE) quantifies deviations between predicted confidence and empirical frequency across subgroups; Counterfactual Invariance Score (CIS) evaluates the agreement of outputs for (x,x^{cf}) when legally relevant causal paths are held constant. The evaluation protocol specifies two channels, attribute-visible and attribute-hidden, to probe robustness under observable and unobservable sensitive features, and conducts paired tests on counterfactual instances to compare interventions including reweighting, temperature recalibration, groupwise recalibration, and invariance regularization. All metrics report both macro and micro averages, with separate breakdowns for long-tail charge types and priority subgroups. The data split, random seeds, retrieval-corpus version, and schema hash are released alongside experiment cards to support third-party verification and apples-to-apples comparisons.

4. Results

4.1. Semantic parsing performance

As shown in Figure 1, the proposed method delivers consistent gains over the baseline on the 2019–2024 corpus: Slot-F1 improves from 84.2 to 88.9 (+4.7, +5.6%), Span-F1 from 81.7 to 86.3 (+4.6, +5.6%), Graph-F1 from 76.5 to 82.4 (+5.9, +7.7%), and Structural Validity Rate (SVR) from 89.1 to 96.7 (+7.6, +8.5%). The larger margins on structure-sensitive metrics (Graph-F1, SVR) indicate that ontology constraints and structural validation primarily correct relation-edge errors and field omissions, while concurrent gains on Slot/Span suggest retrieval evidence strengthens element recall and boundary localization. Stratified aggregation reveals higher relative gains in low-resource categories (juvenile, minor violent offenses), and residual errors shift from slot omissions to tail

DOI: 10.54254/2755-2721/2026.30609

cases of relation ambiguity, implying that long event chains and cross-paragraph evidence alignment should be further constrained. Overall, the simultaneous rise of structural validity and relation-level F1 corroborates that improvements stem from enhanced event-chain integrity and rationale alignment rather than mere accumulation of slot predictions.

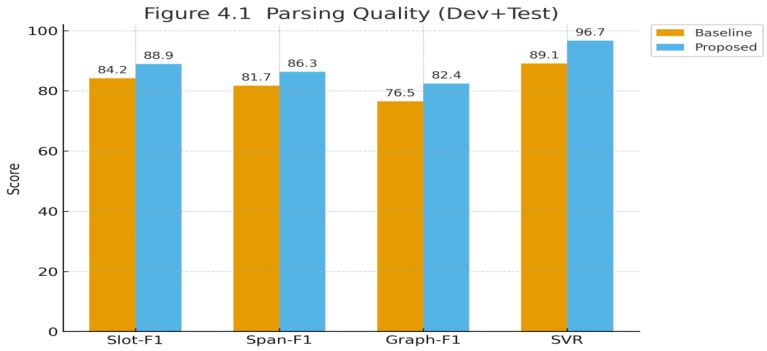

Figure 1. Parsing quality (Dev+Test)

4.2. Bias calibration and downstream recommendations

As shown in Table 1, calibration markedly reduces subgroup disparities under both attribute-visible and attribute-hidden protocols. In the visible protocol, COG decreases from 6.8% to 2.1% (−4.7pp, −69%), EOD from 7.5% to 2.9% (−4.6pp, −61%), GCE from 5.6% to 2.4% (−3.2pp, −57%), while CIS rises from 0.78 to 0.91 (+0.13). In the hidden protocol, COG drops from 5.1% to 1.8% (−65%), EOD from 6.2% to 2.5% (−60%), GCE from 4.7% to 2.0% (−57%), and CIS increases from 0.81 to 0.90 (+0.09). The convergence patterns are consistent across protocols, with slightly larger reductions when attributes are visible, indicating that reweighting and groupwise recalibration more effectively target observable subgroup gaps. The simultaneous increase in CIS and decrease in GCE shows reduced sensitivity to non-causal attributes and better confidence–frequency alignment, without sacrificing parsing accuracy (cf. Figure 4.1 improvements in F1 and SVR). Residual gaps concentrate in long-tail charge types and cross-domain texts with proxy features, suggesting future work on expanding counterfactual coverage and refining subgroup-specific thresholds to compress tail asymmetries.

Table 1. Bias calibration evaluation

Protocol	Stage	COG (%)	EOD (%)	GCE (%)	CIS
Attribute-Visible	Raw	6.8	7.5	5.6	0.78
Attribute-Visible	Calibrated	2.1	2.9	2.4	0.91
Attribute-Hidden	Raw	5.1	6.2	4.7	0.81
Attribute-Hidden	Calibrated	1.8	2.5	2.0	0.90

5. Discussion

Findings indicate that ontology constraints and structural validation are decisive for relation-edge correctness and field completeness, evidenced by larger gains in Graph-F1 and SVR relative to Slot/Span. Retrieval-augmented evidence attenuates cross-domain expression drift, yet long event chains and cross-paragraph linkages remain the dominant error sources. Bias calibration converges under both attribute-visible and attribute-hidden protocols; the rise in CIS alongside the decline in

DOI: 10.54254/2755-2721/2026.30609

GCE shows suppressed responsiveness to non-causal sensitive attributes and improved confidence–frequency alignment. Residual asymmetries concentrate in long-tail charge types and proxy-feature contexts, suggesting the need to broaden counterfactual coverage and refine subgroup-specific thresholds. Limitations include annotation sensitivity to ontology granularity, jurisdictional heterogeneity that bounds transferability, and compliance constraints around sensitive-attribute collection and de-identification. Human–AI co-review and longitudinal monitoring are indispensable safeguards.

6. Conclusion

This study presents an integrated framework for semantic parsing and bias calibration tailored to restorative justice. The approach enforces ontology-driven constrained decoding and explicit structural validation over an "events–roles–measures–outcomes" schema, stabilized by retrieval evidence, and couples counterfactual generation, cost-sensitive reweighting, and groupwise/temperature recalibration into a reproducible RJ-BCB evaluation loop. Empirically, the method improves Graph-F1 and SVR without sacrificing parsing accuracy, while substantially reducing COG/EOD and increasing CIS with better inter-group calibration. For deployment, we recommend routing structured outputs directly into evidence-chain visualization and review workflows, with subgroup-level thresholds and audit dashboards for continuous oversight. Future work will strengthen constraints for procedural pragmatics and long-chain causality, expand multimodal evidence integration, and study robust recalibration under distributional shift.

References

[1] Chalkidis, Ilias, et al. "LexGLUE: A benchmark dataset for legal language understanding in English." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.
[2] Guha, Neel, et al. "Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models." Advances in neural information processing systems 36 (2023): 44123-44279.
[3] Niklaus, Joel, et al. "Multilegalpile: A 689gb multilingual legal corpus." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024.
[4] Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Stürmer, M., & Chalkidis, I. (2023). Lextreme: A multi-lingual and multi-task benchmark for the legal domain. arXiv preprint arXiv: 2301.13126.
[5] Louis, Antoine, and Gerasimos Spanakis. "A statutory article retrieval dataset in French." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.
[6] Rabelo, Juliano, et al. "Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021." The Review of Socionetwork Strategies 16.1 (2022): 111-133.
[7] Silva Filho, Telmo, et al. "Classifier calibration: a survey on how to assess and improve predicted class probabilities." Machine Learning 112.9 (2023): 3211-3260.
[8] Jiang, Zhengping, Anqi Liu, and Benjamnin Van Durme. "Addressing the binning problem in calibration assessment through scalar annotations." Transactions of the Association for Computational Linguistics 12 (2024): 120-136.
[9] Ulmer, Dennis, et al. "Calibrating large language models using their generations only." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024.
[10] Hansen, Dutch, et al. "When is multicalibration post-processing necessary?." Advances in Neural Information Processing Systems 37 (2024): 38383-38455.
[11] Perez-Lebel, Alexandre, Marine Le Morvan, and Gaël Varoquaux. "Beyond calibration: estimating the grouping loss of modern neural networks." arXiv preprint arXiv: 2210.16315 (2022).