Assessing Legal Argumentation in Generated Judgments: A Toulmin-Based Evaluation Framework

Cor Steging
Department of Information and Computing Sciences,
Utrecht University, the Netherlands

Tadeusz Zbiegień
Department of Legal Theory,
Jagiellonian University, Poland

Abstract

Legal judgments require sound legal reasoning in order for justice to occur. Existing work on automated judgment prediction has largely focused on predicting the final decision, while ignoring the quality of the underlying legal argumentation. Existing semantic similarity metrics evaluate the semantic similarity of generated texts, but it remains unclear whether they reflect the legal correctness of judicial reasoning. This paper makes two contributions: first, a novel framework for evaluating the legal argumentation of generated judgments based on Toulmin models; second, an application of this framework to a set of generated judgments of appellate negligence cases, and a comparison with contemporary semantic similarity metrics. Our results show that, despite achieving competitive performance on outcome prediction, the evaluated model frequently produces legally deficient reasoning. Moreover, we find that contemporary semantic similarity metrics do not capture the same judicial argumentation qualities as our framework.

1 Introduction

AI is increasingly being applied in the legal domain, including for tasks such as legal prediction and automated decision support. However, predicting legal outcomes alone is not sufficient, as legal decisions require sound argumentation that justifies the final ruling. Recent advances in large language models (LLMs) make it possible not only to predict legal outcomes, but also to generate the underlying reasoning supporting those decisions.

While outcome prediction is relatively straightforward to evaluate, assessing the quality of legal reasoning at scale remains challenging. Although a range of semantic similarity metrics exist for measuring textual similarity, it is unclear to what extent these metrics reflect legal correctness or soundness of argumentation.

In this paper, we evaluate LLMs on both outcome prediction and the quality of the reasoning underlying their decisions. In particular, we investigate whether contemporary semantic similarity metrics serve as reliable proxies to assess legal argumentation by comparing them against a novel expert-based annotation framework grounded in Toulmin models.

2 Background

Related work frames judicial-reasoning generation as a legal natural language generation task (Ye et al., 2018; Yue et al., 2024; Li et al., 2024; Nigam et al., 2024). Argument mining aims at extracting structured arguments from unstructured text (Lippi and Torroni, 2016), with legal argument mining showing in particular that judicial reasoning requires domain-specific qualities (Habernal et al., 2024). In this study, we want to approach the evaluation of generated legal argumentation based on Toulmin's model of argumentation (Toulmin, 1958). Toulmin models, as well as other argument schemes, have played a key role in the AI & Law community due to their effectiveness in capturing legal reasoning, resulting in a long-standing tradition of using Toulmin models to represent legal argumentation (Prakken, 2006; Atkinson and Bench-Capon, 2021)

In combination with LLMs, Toulmin models have previously been used to prompt LLMs to explicate the argumentative structure of informal texts, extracting claim-reason-warrant triples (Gupta et al., 2024). The TRACE metric (Kim and Yang, 2026) applies a similar theoretical basis to LLM reasoning traces, using a fine-tuned classifier to score both the validity of individual Toulmin and metacognitive states and the coherence of transitions between them. In contrast, we derive scoring rubrics from Toulmin models of ground-truth judgments and have legal experts apply them to

Proceedings of the 19th International Natural Language Generation Conference: Short Papers, pages 50–58
October 17-21, 2026 ©2026 Association for Computational Linguistics

John Doe vs Jane Doe
Case ID: 190
Facts: On May 5, 2024, Jane Doe was...
Procedural History: The trial court assessed no negligence against plaintiff...
Application of Law to Facts: Defendant correctly argues that...
Relevant Precedents: Johson vs. Williams (2023)...
Outcome: Reverse

Figure 1: Fictional example case

assess the legal soundness of LLM-generated judgments. Argumentation in general has also been used in combination with LLMs, such as evaluating whether LLMs can understand argument schemes (Bezou-Vrakatseli et al., 2025).

3 Method

This section describes our dataset, the novel Toulmin-based argumentation assessment framework, the semantic similarity metrics we compare it against, and the experimental setup.

3.1 Dataset

We use the AnnoCaseLaw dataset (Sesodia et al., 2025), which consists of 394 annotated U.S. appellate negligence cases. Each case originates from a negligence dispute that was decided by a trial court and subsequently appealed to a higher court. The dataset decomposes each appellate opinion into five semantically distinct components: the facts, procedural history, relevant precedents, application of the law to the facts, and the appellate outcome (see Figure 1 and a description of each component in Appendix A.1).

In our study, we provide a large language model with the facts and procedural history and prompt it to predict both the final outcome and the application of law to the facts (ALF). This setup allows us to evaluate not only the correctness of the final decision but also the quality of the legal arguments produced by the model.

3.2 Toulmin Argumentation Assessment Framework

To evaluate the legal quality of the generated ALF, we introduce an expert annotation framework based on Toulmin's model of argumentation (Toulmin, 1958). Toulmin's model distinguishes six components of an argument: a *claim*, the conclusion advanced by the argument; *grounds*, the facts or evidence offered in support of the claim; a *warrant*, the inferential rule that connects the grounds to the claim; *backing*, the authority or further support for the warrant; a *qualifier*, the degree or scope of the claim; and a *rebuttal*, an exception, counterargument, or condition under which the claim would not hold.

The annotation proceeds in two stages. In the first stage, legal experts create a grading rubric based on Toulmin models of the ground-truth judgements. For each case, the content that instantiates each Toulmin components is identified and recorded. An example of such a model can be found in Appendix A.2. Non-argumentative connective or purely narrative material is not scored unless it performs one of the Toulmin functions above. Because not every judgment contains all six Toulmin components, components that are not represented are marked as not applicable.

In the second phase, the generated judgments are assessed against the case-specific Toulmin rubric by annotators. Each applicable component receives a score of 0 (no overlap), 1 (partial overlap), or 2 (full overlap). The final Toulmin score for case i is then computed as:

$$\text{ToulminScore}(i) = \frac{10}{2|C_i|} \times \sum_{c \in C_i} s_{ic}$$

where C_i is the set of applicable Toulmin components for case i, and $s_{ic} \in \{0, 1, 2\}$ is the score assigned to category c. This normalizes all cases on a 0–10 scale while avoiding penalties for categories that are not present in the reference judgment.

3.3 Semantic similarity metrics

We compare the Toulmin scores against 27 automatic semantic similarity metrics, following (Huidrom et al., 2025), spanning several families (edit- and set-based, embedding-based, overlap-based, learned regression-based, and LLM-as-a-judge; full list in Appendix A.3). For LLM-as-a-Judge, we use GPT-5.5, Claude Sonnet 5, and Gemini 3.1 Flash Lite, reporting both individual judgments and ensemble performance, with prompts adapted from Zheng et al. (2023) (Appendix A.4).

3.4 Experimental set-up

We provide the facts and procedural history of all 394 cases to a large language model and obtain both the predicted case outcome and the generated

Table 1: Comparison of class-weighted F1 scores for outcome prediction.

Model	Weighted F1
Current work	
GPT-5.5	0.60
(Sesodia et al., 2025)	
GPT-4o	0.47
o3-mini	0.42
Gemini 1.5 Pro	0.48
Random baseline	0.27

application of law to the facts (ALF). We first evaluate the performance of the predicted outcomes. Next, we compute a range of automatic semantic similarity metrics between the generated ALF and the ground-truth ALF.

For a randomly subsample ($n = 100$, seed = 42), we construct Toulmin models of the ground-truth ALF and assign Toulmin scores.[1] We then compare the automatic similarity-based evaluations with these manual assessments using Pearson's correlation coefficient (r).

For the experiments, we use GPT-5.5 from OpenAI, which is, at the time of writing, among the most widely used flagship models [2]. We set the reasoning effort parameter [3] to medium.

4 Results

We report on both the overall classification performance of the model, as well as on the

4.1 Overall performance

We compute the class-weighted F1 score for the GPT-5.5 model, obtaining a score of 0.60. In Table 1, we report this result alongside the random baseline and the performance of earlier models from prior work. Table 2 presents the corresponding confusion matrix, providing a more detailed view of the model's prediction behavior.

4.2 Evaluating argumentation

The automated semantic assessment metrics vary considerably, though closely related variants of the same metric can correlate strongly, as shown for the LLM-as-a-Judge scores in Table 3. The full correlation matrix is provided in Appendix A.5.

[1]The results, Toulmin models, and Toulmin scores can be found at https://github.com/CorSteging/INLG2026

[2]https://platform.openai.com/docs/models, accessed: June 2026

[3]https://platform.openai.com/docs/guides/reasoning, accessed: June 2026

Table 2: Confusion matrix for outcome prediction.

	Predicted			
True	**Affirm**	**Mixed**	**Reverse**	**Support**
Affirm	128	11	44	183
Mixed	23	16	18	57
Reverse	46	12	93	151

Table 4 shows an overview of the Toulmin scores for outcomes predicted correctly versus incorrectly. Inter-annotator agreement was assessed on a subset of 10 cases independently annotated by a third legal expert, to establish the reliability of the Toulmin-based scores; Krippendorff's alpha with ordinal weighting for the final Toulmin score and individual component scores is shown in Table 6. Pearson's r and corresponding p-values comparing the automated semantic similarity metrics against the Toulmin score are shown in Table 5.

5 Discussion

We here discuss the overall performance of the system, compare the Toulmin scores against the automated similarity metrics, and reflect on the limitations of the study and points for future research.

5.1 Overall performance

Table 1 shows GPT-5.5 outperforms both the random baseline and prior models, likely reflecting recent LLM advances, though a weighted F1 of 0.60 indicates room for improvement. The confusion matrix (Table 2) shows the model performs best on majority classes, correctly predicting 70% of affirm and 61.6% of reverse decisions, but only 28.1% of mixed decisions. Most errors involve confusion between affirm and reverse, suggesting a difficulty distinguishing these classes. Overall, the results indicate competitive performance.

5.2 Annotation process

Krippendorff's alpha (Table 6) indicates strong agreement between annotators, on both the final Toulmin score and individual components, despite the small overlap sample ($n = 10$). Backing showed comparatively lower, though still substantial, agreement, likely reflecting greater subjectivity in assessing supporting justification for a warrant.

5.3 Evaluating argumentation

Table 3 shows the three LLM judges agree strongly in pairwise ranking ($r = 0.82$–0.92), but differ

Table 3: Descriptive statistics and pairwise Pearson correlations of the LLM-as-a-Judge scores.

	Statistics					Pearson's r		
Model	**Mean**	**Std.**	**Median**	**Min**	**Max**	**GPT**	**Gemini**	**Claude**
Gemini 3.1 Flash Lite	7.107	3.484	9	1	10	0.915	1.000	0.824
GPT-5.5	6.010	2.740	7	1	9	1.000	0.915	0.883
Claude Sonnet 5	4.802	2.615	5	1	9	0.883	0.824	1.000

Table 4: Toulmin scores overview.

Prediction	n	**Mean**	**Std**	**Range**
Correct	57	6.01	1.97	1.25–10
Incorrect	43	6.22	2.50	0–10
Total	100	6.10	2.20	0–10

Table 5: Pearson correlation between automated metrics and expert-annotated Toulmin scores.

Metric	**Pearson** r	p**-value**
ROUGE-L (Recall)	0.203	0.043*
ROUGE-1 (Recall)	0.197	0.049*
Levenshtein similarity	-0.169	0.093
Universal Sentence Encoder	0.158	0.115
Levenshtein distance	0.130	0.199
BLEURT	-0.116	0.250
Sentence-BERT (Euclidean)	-0.116	0.252
ROUGE-2 (Precision)	-0.111	0.271
ROUGE-1 (Precision)	-0.110	0.277
Sentence-BERT (Manhattan)	-0.109	0.282
METEOR	-0.108	0.285
ROUGE-2 (Recall)	0.107	0.291
Sentence-BERT (Cosine)	-0.105	0.296
Sentence-BERT (Dot Product)	-0.105	0.296
ROUGE-L (Precision)	-0.095	0.347
SacreBLEU	-0.092	0.360
Jaccard distance	0.063	0.534
Jaccard similarity	-0.063	0.534
ROUGE-2 (F1)	-0.060	0.551
Dice coefficient	-0.060	0.555
ROUGE-1 (F1)	-0.058	0.567
BERTScore	-0.052	0.609
GPT-5.5	0.046	0.649
Gemini-3.1 Flash Lite	0.043	0.671
LLM ensemble	0.037	0.715
ROUGE-L (F1)	-0.033	0.744
Claude Sonnet 5	0.013	0.897

Table 6: Krippendorff's alpha between annotators, per Toulmin component and for the derived final score.

Component	α
Claim	0.986
Grounds	0.930
Warrant	0.961
Backing	0.778
Qualifier	1.000
Rebuttal	1.000
Final score	0.993

in scoring tendencies: Gemini is most generous (highest mean/median, only model reaching the maximum score of 10) and most variable, GPT-5.5 scores more conservatively with a narrower spread, and Claude Sonnet 5 consistently assigns the lowest scores.

Table 4 shows Toulmin scores average around 6.1. A Mann-Whitney U test found no significant difference in Toulmin score between correct and incorrect predictions ($U = 1128.0$, $p = 0.4915$), suggesting correct outcome prediction does not necessarily coincide with sound legal argumentation, and vice versa: a case can reach the wrong outcome while still being reasoned relatively well, or reach the right outcome via comparatively weak reasoning.

Comparing Toulmin scores to automated semantic similarity metrics (Table 5) shows uniformly low correlations, clustering near $r = 0$. Two metrics yield $p < 0.05$ ($r = 0.197$, $r = 0.203$), but this is close to the number of false positives expected by chance given 27 tests, and neither survives a Bonferroni correction ($\alpha = 0.05/27 \approx 0.0019$). This suggests standard semantic similarity metrics, including LLM-as-a-judge approaches, do not capture the same legal qualities assessed by our Toulmin-based framework.

Based on a qualitative inspection of the reasoning of the LLM-as-a-judge approaches, we discovered that disagreements often stem from differing strictness. LLM-judges tended to accept plausible reasoning, whereas our Toulmin-based rubrics require specific overlap. Conversely, some generated ALFs received relatively high Toulmin scores despite reaching a conclusion opposite to the reference judgment.

5.4 Limitations and future research

Since our current rubrics score each Toulmin component independently, an explicit structural category checking coherence between components and the final outcome could help address this in future work. Future research should therefore focus on the relationships that connect the elements of the

Toulmin model. In particular, it should examine how the individual nodes (claims, data, warrants, backing, etc.) are linked to one another.

Furthermore, future work should evaluate false positives arising when the LLM generates components absent from the reference model (e.g., an unfounded rebuttal), which our current framework does not penalize. While our framework captures legal argumentation qualities that automated measures do not, it remains a manual and lengthy process. Nonetheless, our results show that automated measures cannot currently replace such expert-level evaluation.

Our evaluation is further limited to a single case study (U.S. appellate negligence cases), and the manually labeled Toulmin models are limited in number; expanding this set would strengthen future analysis. Beyond scaling up, two directions stand out: automating Toulmin extraction and rating with LLMs, comparing such annotations against expert-based ones; and, on the generation side, prompting LLMs to explicitly follow a Toulmin-based structure to improve generated legal judgments. The focus of the current study, however, is evaluation of legal argumentation itself.

6 Conclusion

Existing automatic similarity metrics do not effectively capture legal reasoning. We therefore introduce a Toulmin-based framework for assessing the legal quality of generated judgements and arguments. In an experiment on U.S. negligence appellate cases, we show that our Toulmin-based method reliably captures different argumentation qualities than standard automated semantic similarity metrics.

Acknowledgments

This material is produced as part of AlgoSoc, a collaborative 10-year research program on public values in the algorithmic society, which is funded by the Dutch Ministry of Education, Culture and Science (OCW) as part of its Gravitation programme (project number 024.005.017).

The research leading to these results has also received funding from the National Science Centre, Poland, project no. 2025/57/N/HS5/01561, titled "Uncertainty and Argumentation: Decision-Making Under Uncertainty in Legal Disputes."

We also thank İrem Mutlu, LL.M., for her contribution as second annotator.

References

Katie Atkinson and Trevor Bench-Capon. 2021. Argumentation schemes in AI and law. *Argument & Computation*, 12(3):417–434.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2025. Can large language models understand argument schemes? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13666–13681, Vienna, Austria. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. 2024. Harnessing Toulmin's theory for zero-shot argument explication. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10259–10276, Bangkok, Thailand. Association for Computational Linguistics.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker Genannt Döhmann, and Christoph Burchard. 2024. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 32(3):1–38.

Rudali Huidrom, Michela Lorandi, Simon Mille, Craig Thomson, and Anya Belz. 2025. Assessing semantic consistency in Data-to-Text generation: A meta-evaluation of textual, semantic and model-based metrics. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 98–107, Hanoi, Vietnam. Association for Computational Linguistics.

Yundong Kim and Heyoung Yang. 2026. TRACE: Toulmin-based Reasoning Assessment through Constructive Elements for LLM CoT Evaluation. *Preprint*, arXiv:2605.29656.

Ang Li, Yiquan Wu, Yifei Liu, Kun Kuang, Fei Wu, and Ming Cai. 2024. Enhancing Court View Generation with Knowledge Injection and Guidance. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5896–5906, Torino, Italia. ELRA and ICCL.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2).

Shubham Kumar Nigam, Aniket Deroy, Subhankar Maity, and Arnab Bhattacharya. 2024. Rethinking legal judgement prediction in a realistic scenario in the era of large language models. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 61–80, Miami, FL, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Henry Prakken. 2006. *Artificial Intelligence & Law, Logic and Argument Schemes*, pages 231–245. Springer Netherlands, Dordrecht.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Magnus Sesodia, Alina Petrova, John Armour, Thomas Lukasiewicz, Oana-Maria Camburu, Puneet K. Dokania, Philip Torr, and Christian Schroeder de Witt. 2025. AnnoCaseLaw: A richly-annotated dataset for benchmarking explainable legal judgment prediction. *Preprint*, arXiv:2503.00128.

Stephen Toulmin. 1958. *The Uses of Argument*. Cambridge University Press, Cambridge, England.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.

Linan Yue, Qi Liu, Lili Zhao, Li Wang, Weibo Gao, and Yanqing An. 2024. Event grounded criminal court view generation with cooperative (large) language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 2221–2230. ACM.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

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 Appendix

A.1 Dataset

Cases in our dataset consists of five distinct parts, as seen in Figure 1. We describe each of these in more detail below.

The facts describe the events underlying the dispute before litigation commenced, including any stipulated or otherwise undisputed facts. The procedural history summarizes the procedural posture of the case, including the nature of the appeal, the outcome at the lower court, the applicable standard of review, and the lower court's reasoning. Together, these two components constitute the information available to the appellate court prior to its decision.

The remaining three components capture the court's decision-making process. The relevant precedents contain the relevant past cases and legislation used by the court to make its decision. The application of the law to the facts contains the court's legal reasoning, explaining how the governing legal principles and precedents are applied to the facts of the case. In addition to the core reasoning, this component includes contextual statements and judicial cues that support the court's analysis.

Finally, the outcome records the appellate court's decision. Each case is assigned one of three labels: *affirm*, indicating that the lower court's decision is upheld; *reverse*, indicating that the decision is overturned; or *mixed*, indicating that the appellate court affirms some aspects of the decision while reversing others.

Previous work on this dataset has focused on predicting the outcome using LLMs, achieving class-weighted F1 scores of 0.47 (GPT-4o), 0.42 (o3-mini), and 0.48 (Gemini 1.5 Pro), compared to a random baseline of 0.27 (Sesodia et al., 2025). Prior research also investigated the effect of including relevant precedents and the application of law to facts as additional inputs. We exclude both in our study, as the inclusion of relevant precedents did not improve performance, and the application of law to facts corresponds to the reasoning component that our work aims to predict.

A.2 Toulmin models

We show an example of a Toulmin model in Table 7. We annotate the content of each Toulmin component based on the ground truth application of law to facts part of the case.

A.3 Metrics

In this study, we evaluate 26 automated semantic similarity metrics.

We first consider **edit- and set-based metrics**, including Jaccard distance, Levenshtein distance, Jaccard similarity, Levenshtein similarity, and the Dice coefficient. These metrics quantify lexical overlap and edit distance between the generated and reference judgments.

Next, we evaluate **embedding-based metrics**. We use BERTScore (Zhang* et al., 2020), which measures semantic similarity using contextualized token embeddings. We also use Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), which produces dense sentence-level embeddings using a Siamese or triplet network based on BERT. Similarity between sentence embeddings is computed using cosine similarity, dot product, Euclidean distance, and Manhattan distance.

For **overlap-based metrics**, we use SacreBLEU (Post, 2018), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). SacreBLEU provides standardized and reproducible BLEU scores, while ROUGE is evaluated using all nine combinations of ROUGE-1, ROUGE-2, and ROUGE-L with recall, precision, and F1. METEOR complements these metrics by incorporating stemming and synonym matching.

We further evaluate **learned regression-based metrics**. Specifically, we use BLEURT (Sellam et al., 2020) with the BLEURT-20 checkpoint (Pu et al., 2021), and the Universal Sentence Encoder (Cer et al., 2018), for which we compute cosine similarity between sentence embeddings.

Finally, we evaluate the **LLM-as-a-Judge metrics** (Li et al., 2025). Following (Huidrom et al., 2025), we use three state-of-the-art models and report both their individual judgments and their performance as an ensemble. The models are GPT-5.5 (OpenAI), Claude Sonnet 5 (Anthropic), and Gemini 3.1 Flash Lite (Google). The prompts we used are based on those introduced by Zheng et al. (2023) and can be found in Appendix A.4.

A.4 Prompts

For reproducibility purposes, we provide the full prompts in this section. Figure 2 contains the prompts given to model to predict the outcome and application of facts to law for a given case, based on the prompts used in Sesodia et al. (2025). For the LLM-as-a-judge metric, we used the prompt found in Figure 3, which is based on the prompt introduced by Zheng et al. (2023).

A.5 Additional results

The full correlation heatmap that displays the pearson's r between the different semantic similarity metrics can be found in Figure 4. This heatmap is ordered by the families we describe in Section A.3

Claim	Grounds	Warrant	Backing	Qualifier	Rebuttal
We hold that the amended complaint properly alleges a valid cause of action. We hold that the allegation of a breach of Section 60-7B-1.1 which caused injury to plaintiffs states a claim for relief and that claim is not barred by the prospectivity rule stated in *Lopez*.	Appellees here are alleged to have violated a New Mexico statute prohibiting the furnishing of liquor to minors. The allegations of appellants' second amended complaint are that appellees bought or furnished alcoholic beverages for Key, a minor, or permitted him to consume alcoholic beverages in violation of the New Mexico Liquor Control Act.	That law was intended to protect minors and to protect the general public from the risk of injury caused by intoxicated minors. It is well settled that intervening acts of the driver will not relieve the original wrongdoer of liability if those acts are reasonably foreseeable.	Youth and inexperience make misuse of alcohol especially likely, and therefore more foreseeable when the drinker is under the age of twenty-one. Public policy dictates that the minors and the public are entitled to be protected from those who would illegally furnish liquor to underaged persons.	N/A.	N/A.

Table 7: Example Toulmin model based on case 412 (Walker v. Key).

You are a seasoned judge specializing in U.S. civil negligence appeals cases. Analyze the following case details and determine the appropriate appellate outcome: `affirm`, `reverse`, or `mixed` if part of the decision is affirmed and part reversed.

In `application_of_law_to_facts`, explain in a formal judicial style how the governing negligence principles (including duty, breach, causation, and damages where relevant) and the applicable appellate standard of review apply to the material facts and support the outcome.

Case details: `{input}`

Reason step by step and be as thorough as possible.
Respond in strict JSON format with a single key-value pair:

```
{
  "outcome": "affirm" | "reverse" | "mixed",
  "application_of_law_to_facts": "A judicial-style explanation applying the relevant
      negligence law and appellate standard of review to the facts, explaining why the
      outcome is affirm, reverse, or mixed."
}
```

Example output:

```
{
  "outcome": "affirm",
  "application_of_law_to_facts": "Claimant failed to show that Respondent had a duty to lock
      the fire extinguishers in its penal institution. There is no rule or regulation to that
       effect. In addition, Claimant failed to show that Respondent had notice of a dangerous
       condition. In fact, testimony showed that no one had used a fire extinguisher as a
      weapon at the Centralia Correctional Center in at least the past 15 years. Respondent
      has decided that fire safety warranted keeping fire extinguishers unlocked. It appears
      that Respondent acted reasonably and in a manner consistent with accepted prison
      practices."
}
```

Figure 2: Prompt for generating outcome and ALF

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant. Your evaluation should consider semantic and legal similarity. You will be given a reference answer and the assistant's answer.
Begin your evaluation by comparing the assistant's answer with the reference answer. Identify and correct any mistakes. Be as objective as possible.

Reference answer: `{reference_answer}`

Assistant's answer: `{assistant_answer}`

You must rate the response on a scale of 1 to 10 in strict JSON format with exactly two key-value pairs:

```
{
  "rating": 1-10,
  "reasoning": "Concise summary of the evaluation and justification for the rating."
}
```

Figure 3: Prompt for LLM as a judge

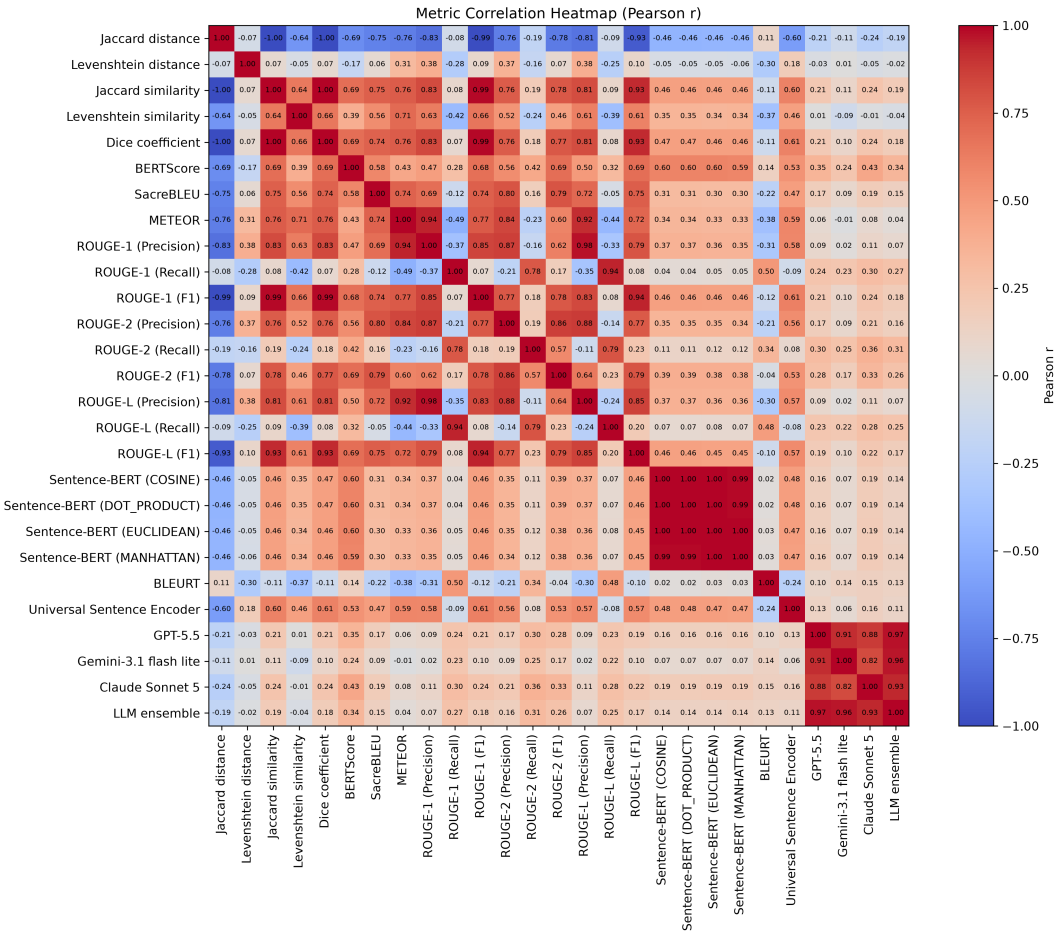


Figure 4: Correlation between the automated semantic similarity metrics (Pearson's r)