

Two Lawyers, Three Opinions: Can Output-Based LLM Uncertainty Find the Cases Where Experts Disagree?

Laura Braun^{1,2}, Matthias Aßenmacher^{2,3}, Martin Hohenadler¹, Göran Kauermann^{2,3}

¹Central Office for Information Technology in the Security Sector (ZITiS)

²Department of Statistics, LMU Munich ³Munich Center for Machine Learning (MCML)

Correspondence: Laura.Braun@stat.uni-muenchen.de

Abstract

Legal natural language processing (NLP) tasks often lack a single objective ground truth, as reflected in label variation across experts’ annotations. We test whether two uncertainty signals based on outputs of large language models (LLMs), sampling entropy and verbalized probability, can identify posts where legal experts disagree. On 1,564 German social media posts annotated by three legal experts for punishability under the German Criminal Code, we evaluate four open-weight LLMs, with reasoning either enabled or disabled, and observe a drop in F1 score on contested posts. Sampling distributions collapse to near-deterministic judgments in most conditions, while verbalized probabilities track expert disagreement in aggregate and support deferral of model errors. Neither signal, however, reliably identifies individual contested posts: both detect the model’s own errors (AUROC up to 0.83) better than expert disagreement (up to 0.76), and which signal is more informative depends strongly on model and activation of reasoning mode.

1 Introduction

Most contemporary NLP systems are trained and evaluated on the assumption that every input has a single correct label (Aroyo and Welty, 2015; Uma et al., 2021). In legal domains, where law is interpretive rather than empirical, this assumption breaks down: Whether a provision of the German Criminal Code (StGB) applies to a given statement is often a matter of legal judgment rather than fact, and detecting punishable online content inherits this ambiguity. On our data, disagreement persists even among three legal experts, who diverge on 27% of posts (Figure 6). Such variation has several reasons (Plank, 2022; Xu et al., 2024): the post itself may be ambiguous, i.e. borderline under the statute, or the experts may vary in how they apply the criteria. Three single judgments per post found item ambiguity with annotator variation, so

we call a post contested without claiming a source. Following Plank (2022), we treat this variation as information that aggregation to a single gold label discards, and use expert disagreement as the signal against which we benchmark model uncertainty.

We use data with annotations by legal experts from the KISTRA project (Lenz, 2024), conducted with the participation of the German Federal Criminal Police Office (BKA). On this kind of task, model performance lags well behind expert judgment (Ludwig et al., 2025), yet the scores are still measured against a single aggregated label. If even experts cannot settle some cases, that may set the wrong objective, and a more attainable goal would be for models to recognize and flag them.

Contributions. We benchmark two output-based uncertainty signals, sampling entropy and verbalized probability. Neither of these requires access to model internals or additional training, which makes them a practical starting point for uncertainty quantification in applied legal NLP. Prior work using those signals (see Section 2) focuses on disagreement among crowd workers on subjective tasks. We ask (RQ1) whether the signals identify contested posts, and (RQ2) whether they are reliable enough to defer uncertain posts to human review.

2 Related Work

Human label variation. Annotator disagreement is increasingly treated as signal rather than noise. Pavlick and Kwiatkowski (2019) show that disagreement in natural language inference persists as more judgments are collected; Nie et al. (2020) release per-item human distributions as soft labels to evaluate against; and Davani et al. (2022) model annotators individually instead of aggregating. This work studies non-expert annotators on subjective tasks. Our annotators are legal experts applying statutory criteria rather than lay annotators reporting personal views.

Output-based Uncertainty. Whether a model aligns with this human variation can be assessed without access to its internals: verbalized confidence has the model state its uncertainty in words (Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024), while sampling-based estimates repeat the query and read uncertainty off the spread of the resulting judgments (Lee et al., 2023). Both yield a per-item score for ranking posts by certainty and deferring the most uncertain to human review (El-Yaniv and Wiener, 2010). Lee et al. (2023) find that the sampled distribution fails to match the human one on NLI, and that alignment degrades further on high-disagreement items.

Uncertainty in Legal NLP. Ludwig et al. (2025) classify punishable incitement to hatred under § 130 StGB and report a persistent gap between LLMs and legal experts regardless of legal conditioning, weakest on the interpretive elements of the offense. Xu et al. (2024) calibrate case-outcome classifiers (fine-tuned encoders) against judges’ split votes at the European Court of Human Rights: the models separate split from unanimous cases but align only weakly with the vote distribution itself. Closest to our work, Ni et al. (2026) compare the two signals on crowd-annotated data and find that verbalized distributions capture disagreement better while sampling predicts the majority label better.

3 Experimental Setup

Data. The dataset comprises 1,564 social media posts retrieved from the KISTRA corpus (Lenz, 2024), each annotated by three legal experts for punishability under the German Criminal Code and, where applicable, the specific section.¹ A detailed description of the data is given in Appendix A. By majority vote, 24% of posts are punishable. We call a post contested when the experts split 2–1 (424 posts, 27.1%) and unanimous otherwise.

Inference. Each post was judged based on a fixed, four-step checklist (candidate provisions, statutory criteria, freedom-of-expression balancing, decision), decomposing the assessment into the elements of the offense following Zufall et al. (2022); the full prompt is given in Appendix C. Each model returns a binary judgment for punishability, the applicable section, and an integer (0 to 100) probability for the post being criminal.

¹ §§ 86, 111, 126, 129, 130, 140, 166, 185 ff., or 241 StGB.

Models and Evaluation. We evaluate Qwen3 (8B, 32B) and Gemma 4 (12B, 31B) models with and without reasoning by drawing ten samples per post at the recommended temperatures.² We report precision, recall, and F1 score against the majority vote (overall and separately for contested and unanimous posts). We call a prediction a model error when it deviates from this majority vote, i.e., from the aggregated expert label, which on contested posts rests on a 2–1 split. We further evaluate both uncertainty signals by selective prediction against model errors, by their ability to separate contested from unanimous posts, and against the human soft label, i.e., the share of experts voting punishable. The separation of contested from unanimous posts and the fit to the soft label address RQ1, while selective prediction against model errors addresses RQ2.

Uncertainty Signals. From the ten samples per post we derive two per-item scores. Sampling entropy is the binary entropy of the share of samples judging the post punishable. The verbalized $P(\text{punishable})$ of a post is the mean of the ten reported values on a 0 to 1 scale, and as an uncertainty score we use its closeness to 0.5. Both scores can be used to rank posts from the most to least uncertain, so that the most uncertain can be withheld from automatic prediction, and e.g. prioritized for human review.

4 Results

Classification. Figure 8 shows that the binary F1 score³ against the majority vote ranges from 0.46 to 0.69 across conditions and models. Stratifying as in Figure 1 shows that F1 on contested posts stays between 0.45 and 0.61, but that the gap to unanimous posts is split by model family: For the Gemma models, F1 drops by 0.15 to 0.17 on contested posts, while for the Qwen models, the two subsets are indistinguishable within their confidence intervals. Model errors, therefore, concentrate where experts disagree for Gemma but not for Qwen models. Precision and recall are reported in Appendix B, where the Gemma (Qwen) models tend to favor precision (recall).

²We use the recommended temperatures (0.6 and 0.7) for Qwen3 (Yang et al., 2025) and (1.0) for Gemma 4 (Gemma Team, 2026). As temperature differs between conditions, we compare entropy by AUROC.

³Prediction for a post is the majority of the ten samples; the 0.4–4.2% with a 5/5 split counts as not punishable.

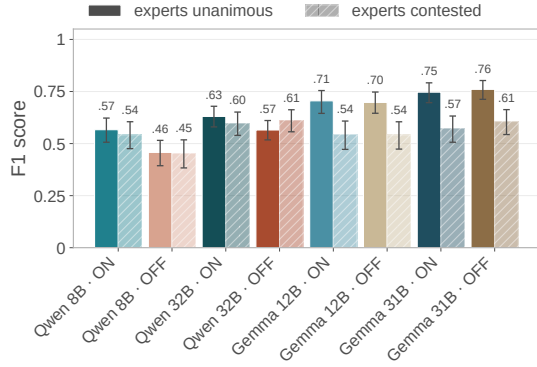


Figure 1: F1 against the majority vote on posts with unanimous expert labels and on contested posts.

Sampling Entropy. Figure 2 shows the distribution of the share of punishable judgments by LLMs, grouped by human soft label. The distributions peak at 0, at 1, or both, but asymmetrically: 52–90% sit entirely at 0 on posts no expert judged punishable, while only 31–63% sit entirely at 1 when all three did, and 4–19% still sit at 0. Without reasoning the collapse is stronger still: up to 95% of posts receive ten identical judgments (Gemma 31B). Ranking items by entropy yields nearly flat risk-coverage curves (Figure 3, top), with error-detection AUROC between 0.62 and 0.71 with reasoning and between 0.53 and 0.60 without. Reasoning helps, but even the best condition removes few errors.

Verbalized Probability. The median verbalized $P(\text{punishable})$ increases monotonically with the human soft label (Figure 4), so the signal does carry information about expert disagreement at the population level. Per item, however, its IQR spans large parts of the unit interval. The Gemma models stay near zero at 0/3–1/3 and rise steeply thereafter, though the 2/3 bucket lands mid-scale (medians 0.50 and 0.67) rather than near one. We measure calibration against the human soft label rather than the majority vote, since Baan et al. (2022) show that calibration to an aggregated label is problematic when annotators genuinely disagree. The verbalized value is also unstable across the ten samples of the same item (Appendix B, Figure 9). As a selective prediction signal, however, it clearly outperforms sampling entropy. Ranking items by closeness of the verbalized probability to 0.5 yields informative risk-coverage curves (Figure 3, bottom), with error-detection AUROC between 0.73 and 0.83 with reasoning and between 0.65 and 0.81

without.

Comparing the Signals. The two signals are not interchangeable. Their per-item rank correlation ranges from 0.80 (Qwen3 32B with reasoning) to -0.14 (Gemma 31B without reasoning) (Figure 10), and reasoning strengthens the coupling for every model. As an estimate of the human soft label, verbalized probability is the more informative signal. It achieves a lower Brier score and Jensen-Shannon divergence than the sampling distribution in all eight conditions (Appendix B, Figure 11). Detecting expert disagreement itself, the target of RQ1, is harder for both signals (Figure 5). Discriminating contested from unanimous posts, verbalized uncertainty reaches AUROC values between 0.63 and 0.76 while sampling entropy stays between 0.51 and 0.65, both well below the best error-detection AUROC of 0.83. The picture is the same under AUPRC, which accounts for the 27% share of contested posts: verbalized uncertainty reaches 0.33 to 0.46 and sampling entropy 0.28 to 0.37, against 0.271 for a random ranking. Stratifying by majority class (Appendix B, Figure 7) shows where this separation comes from. Verbalized uncertainty distinguishes 1/3 from 0/3 posts about as well as in the pooled setting, while sampling entropy is near chance in most conditions; on majority-positive posts both signals separate 2/3 from 3/3 close to chance. Across both metrics and both strata, the ordering of the two signals is stable, but neither truly reaches the level at which contestedness could be read off a single post.

5 Discussion and Outlook

Neither signal answers our questions cleanly. For RQ1, the alignment with expert disagreement exists but is weak. Sampling entropy is nearly silent: the judgment is binary, the model commits to it, and repeated samples of the same post rarely diverge. The collapse is directional: sampling generally reproduces each family’s precision–recall asymmetry rather than the experts’ split, which is why entropy is uninformative about contestedness even where it is confidently wrong. Verbalized probability instead recovers the ordering of contested and unanimous posts in aggregate, yet with a moderate AUROC it cannot reliably identify individual contested posts. Regarding RQ2, we conclude that neither signal works as a disagreement router. What verbalized probability does support is error-based deferral. Ranking posts by closeness

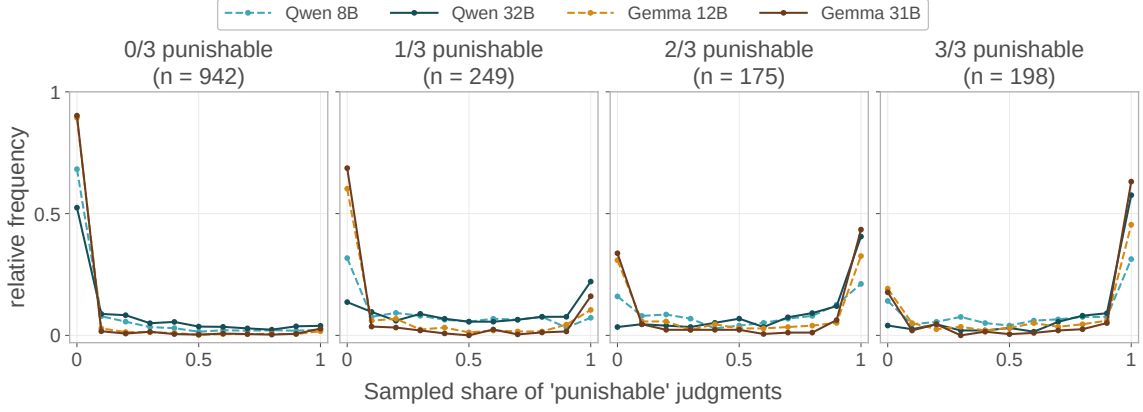


Figure 2: Empirical distribution of the sampled share of 'punishable' judgments (10 samples per item) with reasoning enabled, grouped by expert votes (0/3 to 3/3). The share takes only the 11 values $k/10$.

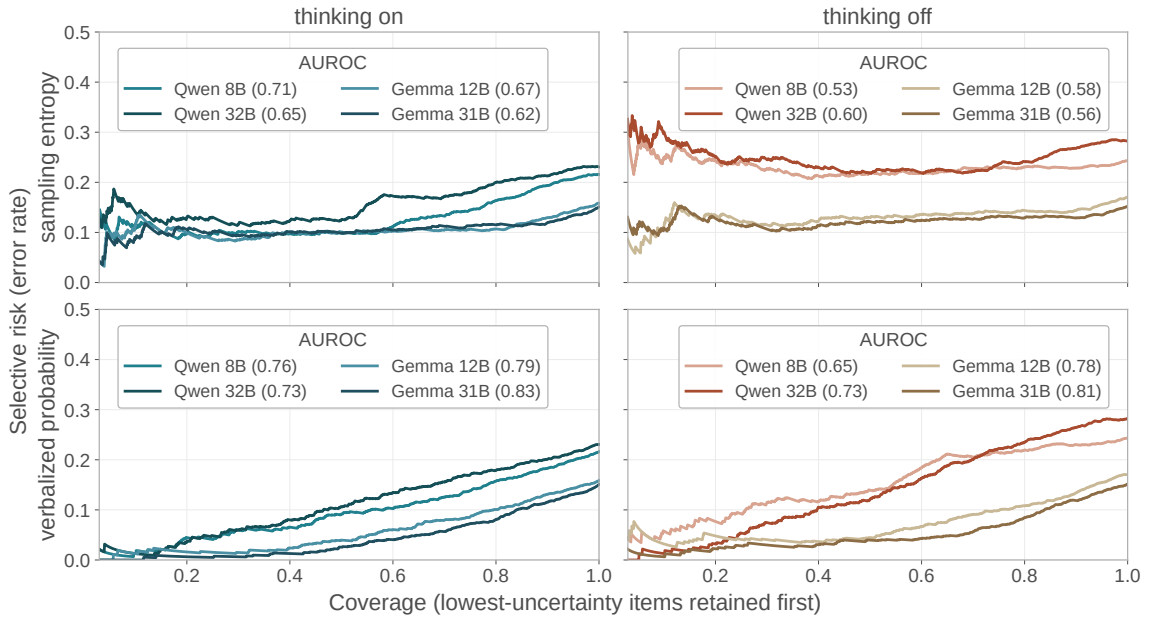


Figure 3: Risk-coverage curves ranking items by sampling entropy (top) and by verbalized uncertainty, i.e. closeness of the verbalized probability to 0.5 (bottom), with AUROC for discriminating errors.

to 0.5 and deferring the most uncertain half sends 71–92% of model errors to review. For Gemma 31B with reasoning, at the upper end of that range, the retained half has an error rate of 2.4%. Part of that is composition, since the retained posts are only 13% punishable against 24% overall, which puts the majority-class baseline on this subset at the same 13%. The model’s error rate is 18% of that baseline. The skew is the cost. Review receives 72% of all punishable posts and 79% of contested posts. The verbalized probability agrees with the models’ own verdict in 98–99.6% of posts and is best read as a graded restatement of that verdict. It tracks how committed the model is, and commitment predicts the models’ own errors far better

than it predicts whether a post is contested.

Two observations qualify this picture. **First**, which signal is informative depends heavily on model family, scale, and reasoning mode. The same recipe that works for one model can be uninformative or even inverted for another, so any deployment must validate its uncertainty signal per model rather than assume transferability. **Second**, although the per-model picture varies, the aggregate ordering is stable. Verbalized probability approximates the human soft label better than the sampling distribution does, as it does for the lay annotators in [Ni et al. \(2026\)](#). Our selective-prediction results extend that ordering to a setting they do not test.

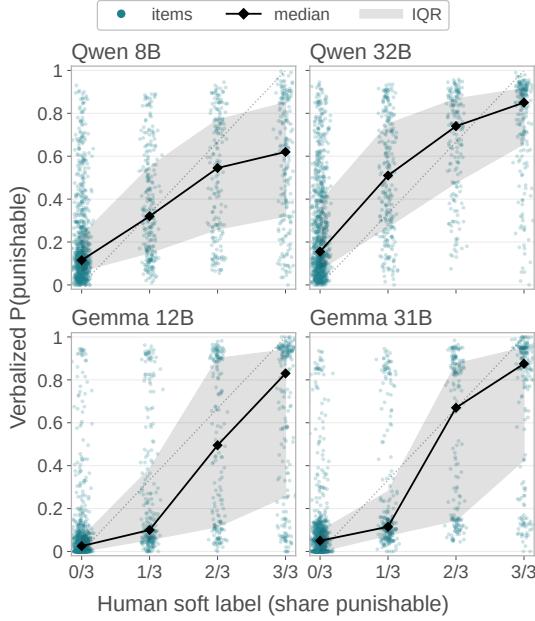


Figure 4: Verbalized $P(\text{punishable})$ with reasoning enabled vs. the share of experts’ votes. The band shows median and IQR per bucket.

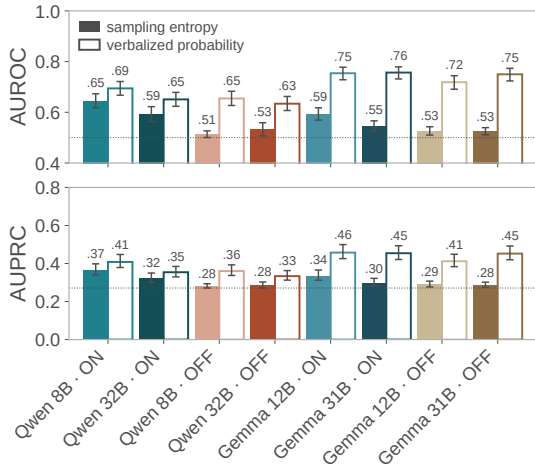


Figure 5: Discriminating contested from unanimous posts. Filled bars rank by sampling entropy, outlined bars by closeness of the verbalized probability to 0.5. Dotted lines mark the value a random ranking attains, 0.5 for AUROC and the share of contested posts (0.271) for AUPRC.

We view these results as a baseline for a domain in which the ground truth itself is contested. Output-based signals tell us more about when a model is likely wrong than about when the post is ambiguous (cf. Baan et al., 2024). Whether that difference is recoverable inside the model through, e.g., probing or disagreement predictors trained on human soft labels remains open, and this benchmark makes it testable.

Limitations

Data access. The raw KISTRA data cannot be released due to data sensitivity and ethical considerations, which limits reproducibility to researchers with equivalent access agreements. Code for inference and evaluation is available at <https://github.com/LauraMBraun/two-lawyers-three-opinions>.

Annotation depth. Three annotations per post may understate true legal ambiguity, and a single round of judgments cannot separate item ambiguity from annotator variation. Separating them would require repeated judgments after a washout period or a larger annotator pool, and recruiting additional qualified legal experts was not feasible within this study.

Comparability confound. Human labels and model signals differ in granularity. Three votes place the human soft label on four levels, and since binary entropy is symmetric around 0.5, ten samples leave sampling entropy with six distinct values, most items at the lowest. Verbalized probability is finer-grained, so part of its advantage may be resolution rather than information.

Prompt sensitivity. We did not vary the prompt systematically, since our goal was evaluating uncertainty quantification rather than optimizing accuracy, though prompt design likely also affects calibration.

Scope of contribution. This paper reports no new UQ method. We instead characterize the usefulness and limits of two easily accessible signals in a domain where the ground truth is itself highly contested.

Ethical considerations

Data sensitivity. The dataset contains posts selected for potential criminal relevance. We therefore report only aggregate statistics. Inference ran only locally on open-weight models and posts were never sent to a third-party API.

Uncertainty is not a substitute for human judgment. Verbalized probability and sampling entropy are behavioral signals, not correctness guarantees, and should not be mistaken for legal certainty.

Cost of errors. This is especially important given the persistent LLM-expert performance gap (Ludwig et al., 2025) and substantial expert disagreement, since false negatives and false positives carry different but equally serious societal costs.

Acknowledgments

We thank Martin Hofer as well as the editors and anonymous reviewers, for their helpful comments and feedback. The Central Office for Information Technology in the Security Sector is funded by the German Federal Ministry of the Interior. The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of any agency of the German government. Matthias Aßenmacher received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 27/1 - 460037581 - BERD@NFDI.

References

- Lora Aroyo and Chris Welty. 2015. [Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation](#). *AI Magazine*, 36(1):15–24.
- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. 2022. [Stop Measuring Calibration When Humans Disagree](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. [Interpreting Predictive Probabilities: Model Confidence or Human Label Variation?](#) In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian’s, Malta. Association for Computational Linguistics.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations](#). *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Ran El-Yaniv and Yair Wiener. 2010. [On the Foundations of Noise-free Selective Classification](#). *Journal of Machine Learning Research*, 11(53):1605–1641.
- Gemma Team. 2026. [Gemma 4 Technical Report](#). *arXiv preprint*. ArXiv:2607.02770 [cs.CL].
- Noah Lee, Na An, and James Thorne. 2023. [Can Large Language Models Capture Dissenting Human Voices?](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.
- Berit Lenz. 2024. [Verbundprojekt: Einsatz von KI zur Früherkennung von Straftaten; Akronym: KISTRA; Teilvorhaben: Anwendung von qualitätsgesicherter KI zur Früherkennung von Straftaten innerhalb des BKA: Abschlussbericht](#). Technical report, Hannover : Technische Informationsbibliothek.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Teaching Models to Express Their Uncertainty in Words](#). ArXiv:2205.14334 [cs.CL].
- Florian Ludwig, Torsten Zesch, and Frederike Zufall. 2025. [Conditioning Large Language Models on Legal Systems? Detecting Punishable Hate Speech](#). In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, Hannover, Germany. HsH Applied Academics.
- Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Miserlis Hoyle, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, and Elliott Ash. 2026. [Can Reasoning Help Large Language Models Capture Human Annotator Disagreement?](#) In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–54, Rabat, Morocco. Association for Computational Linguistics.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. [What Can We Learn from Collective Human Opinions on Natural Language Inference Data?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.
- Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent Disagreements in Human Textual Inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Barbara Plank. 2022. [The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. [Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio.

2021. [Learning from Disagreement: A Survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs](#). In *ICLR 2024 (The Twelfth International Conference on Learning Representations, Vienna, 7–11 May 2024*. arXiv. ArXiv:2306.13063 [cs.CL].
- Shanshan Xu, Santosh T.y.s.s, Oana Ichim, Barbara Plank, and Matthias Grabmair. 2024. [Through the Lens of Split Vote: Exploring Disagreement, Difficulty and Calibration in Legal Case Outcome Classification](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 199–216, Bangkok, Thailand. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 Technical Report](#). *arXiv preprint*. ArXiv:2505.09388 [cs.CL].
- Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2022. A Legal Approach to Hate Speech – Operationalizing the EU’s Legal Framework against the Expression of Hatred as an NLP Task. In *Proceedings of the Natural Legal Language Processing Workshop 2022 (NLLP @ EMNLP 2022)*.

A Detailed Dataset Description

Because the raw data cannot be released, we describe the corpus here as fully as aggregate statistics allow. It consists of 1,564 German-language social media posts drawn from the KISTRA project. Posts are short, with a median length of 157 characters. Each post was independently labeled by three legal experts with the applicable StGB section or "Kein Paragraph" if none applies. The binary label is derived from this. Posts also include auxiliary annotations on ideological background and type of criminality (which were annotated by social and political scientists). By majority vote, 373 posts (23.9%) are punishable. The expert votes split into 942 posts with zero of three votes for punishable, 249 with one, 175 with two, and 198 with three, so 424 posts (27.1%) are contested at the binary level (Figure 6). At the section level, all three experts assigned the same label for 69% of posts, exactly two agreed on 28%, and no two agreed on 3%. At least one expert assigned a section in 622 posts, and at least two distinct positive sections appear in 103 posts. Table 1 reports the distribution of all 4,692 individual section annotations. § 130 (incitement against groups) dominates the positive labels.

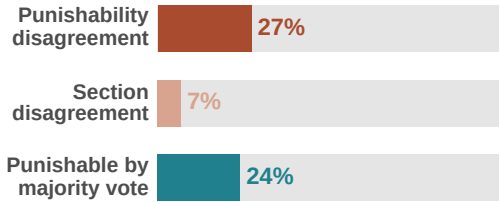


Figure 6: Label distribution across all 1,564 posts. *Section disagreement*: experts named two or more distinct sections (103 posts).

Section	Annotations	Share
Kein Paragraph (none)	3,499	74.6%
§ 130	691	14.7%
§ 185 ff.	194	4.1%
§ 140	158	3.4%
§ 111	57	1.2%
§ 86a	34	0.7%
§ 166	26	0.6%
§ 126	16	0.3%
§ 241	12	0.3%
§ 86	4	0.1%
§ 129	1	0.0%

Table 1: Distribution of the 4,692 individual expert section annotations (three per post).

B Additional Figures

Figures 8 to 11 provide the per-condition detail behind the results reported in the main text.

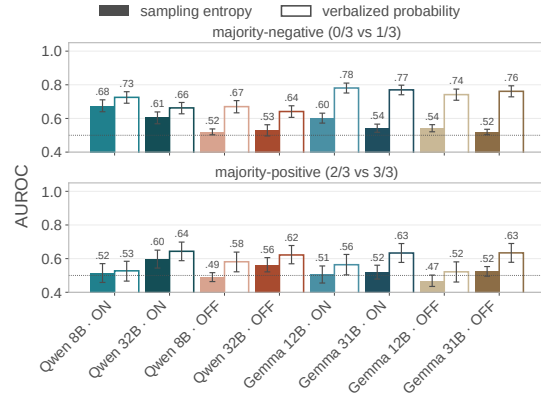


Figure 7: AUROC for discriminating contested from unanimous posts, separately within majority-negative posts (0/3 vs. 1/3, $n = 1,191$) and majority-positive posts (2/3 vs. 3/3, $n = 373$).

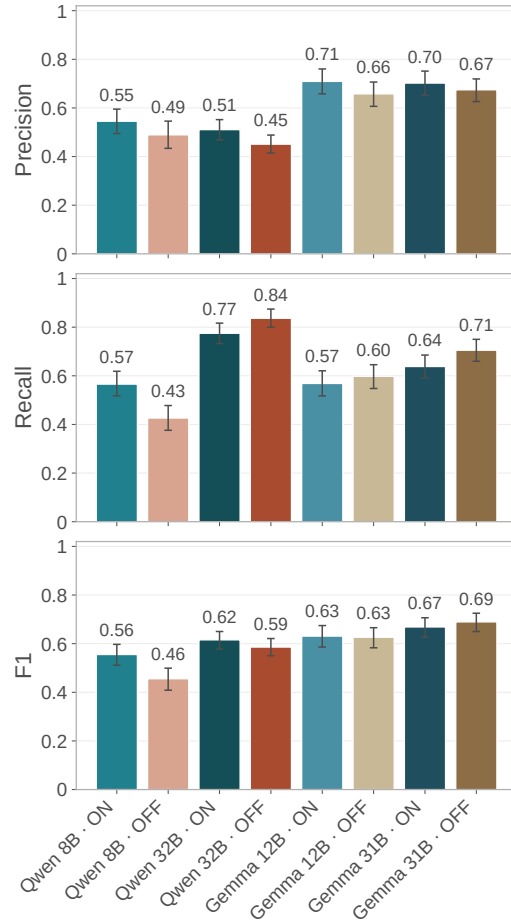


Figure 8: Precision, recall, and F1 for binary punishability against the majority vote across all conditions, with 95% bootstrap confidence intervals.

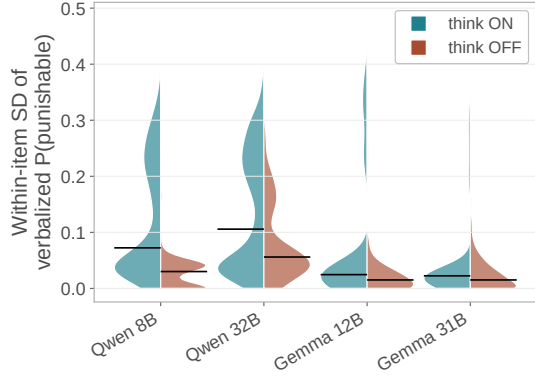


Figure 9: Within-item standard deviation of the verbalized probability across the 10 samples per item (split violins compare reasoning enabled vs. disabled).

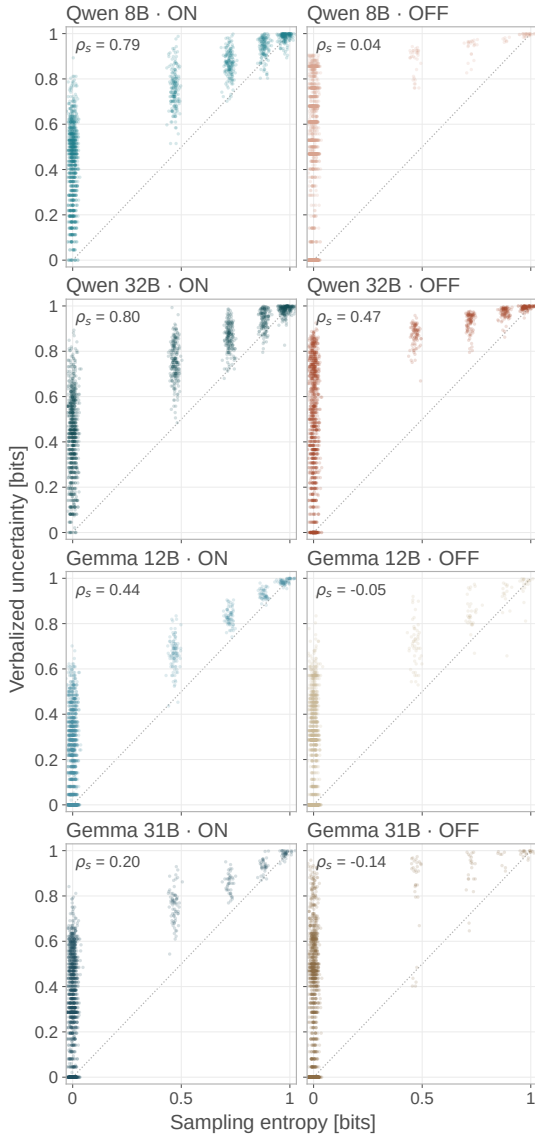


Figure 10: Per-item agreement between the two uncertainty signals, converted to binary entropy, for all models and conditions. ρ_s is the Spearman rank correlation.

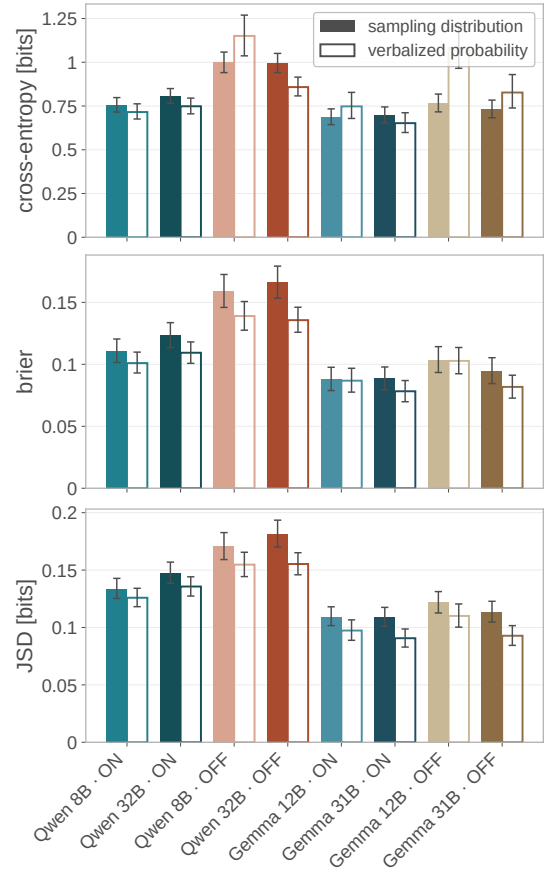


Figure 11: Cross-entropy, Brier score, and Jensen-Shannon divergence between each uncertainty signal and the human soft label (lower is better, 95% bootstrap CIs).

C Prompt

System prompt

You are a trained annotator in a research study on German criminal law. Your task: for a German-language social media post, judge whether the post fulfills the elements of a criminal offense under the German Criminal Code (StGB).

The posts are publicly published social media content collected for research purposes and your judgments are used for scientific annotation only. Work like a briefed research annotator with a fixed checklist. Go through the steps below and decide only at the end. You do not need to write a full legal opinion. What matters is that you apply the statutory criteria and reach a reasoned judgment.

Judge strictly by the statutory criteria. Many posts are offensive, inflammatory, or tasteless without being criminal, and some sound harmless but do fulfill the elements of an offense. Do not let caution in either direction substitute for the legal analysis.

At the end you will report two things:

1. Your judgment (punishable): your single best answer, as a human annotator would give it - true if the post fulfills the elements of an offense, false if not.
2. A probability estimate (p_punishable): an integer from 0 to 100 for how likely it is that the post is in fact criminal. Interpret it as a frequency: among 100 posts where the evidence looks exactly like it does here, in how many would "criminal" turn out to be true? Important: p_punishable is NOT your confidence in your answer.

If you confidently judge a post as not criminal, p_punishable is LOW (close to 0), not high. If you confidently judge it as criminal, p_punishable is high. Only a genuinely uncertain case sits near 50.

Figure 12: System prompt

User prompt template

Assess the following post using the four steps. Think them through internally and at the end, output only the JSON – no intermediate steps in the visible text.

Step 1 – Candidate provisions: briefly check which of the following could plausibly apply: § 130 (Volksverhetzung, incitement against groups), § 185 ff. (Beleidigung, insult), § 111 (Aufforderung zu Straftaten, public incitement to crime), § 140 (Billigung von Straftaten, condoning crimes), or less commonly § 166, § 86a, § 126, § 241, § 129. If none plausibly applies, go straight to Step 4.

Step 2 – Apply the criteria of whichever provision(s) are in play: § 130: Is the statement directed against a group defined by nationality, origin, religion, or ethnicity (or against individuals specifically because of that membership)? And is there, in addition, incitement to hatred, a call for violence, or an attack on human dignity (insulting, maliciously disparaging, or defaming the group)? Mere criticism of a religion, politics, government, or immigration as such is not enough; it requires both the group reference AND one of these acts.

§ 185 ff.: Is contempt expressed toward an identifiable person or a small, clearly delimited group? Cruteness toward "those in power", a party, or a large diffuse group usually isn't enough; a concrete attack on a specific person's honor is required.

§ 111: Is there a public and serious call for a specific crime? A mere wish, sarcasm, or a vague hint is not a call to action.

§ 140: Is a concrete, already-committed serious crime (e.g. a murder or an attack) publicly endorsed?

Step 3 – Check freedom of expression, if a provision from Step 2 could apply: weigh both directions with an open mind. Art. 5 GG (freedom of expression) also covers sharp, polemical, and hurtful criticism – but not everything: a targeted attack on human dignity, the defamation of a specific person, or a serious call for violence falls outside its protection. Decide this concrete case on its own terms, not schematically.

Step 4 – Decide and report: settle on your judgment and your probability estimate, then end your answer with exactly one JSON object and nothing after it, in exactly this shape:

```
{"p_punishable": <integer from 0 to 100>, "punishable": <true or false>,  
"section": "<$ NNN or Kein Paragraph>"}
```

Keep the field names and the literal string "Kein Paragraph" exactly as shown; do not translate it. The angle brackets are type placeholders, not suggested answers. If punishable is false, section is the literal string "Kein Paragraph"; if true, name the single best-fitting provision in "\$ NNN" notation.

Post:

"""

{text}

"""

Figure 13: User prompt template. {text} is replaced by the post under assessment.