Concord: An Agreement-Aware Multi-Adjudication Pipeline for LLM Evaluation

Tyler Bliss*, **Mahit Verma***, **Aila Iyer-Singh, Subrata Biswas, Sheikh Asif Imran, Bashima Islam**
Department of Electrical & Computer Engineering
Worcester Polytechnic Institute
Worcester, MA 01609
{twbliss, mverma, nkiyersingh, sbiswas, simran, bislam}@wpi.edu

Abstract

Evaluating multimodal generations is challenging: human evaluation is costly, and single-model LLM-as-a-judge pipelines can be brittle and provide limited uncertainty signals. We introduce **Concord**, an ensemble-based evaluation pipeline that aggregates discrete judgments from multiple LLM judges and uses inter-judge agreement as a practical uncertainty signal for disagreement-driven triage. We evaluate Concord on **AVSSD** and **SCORE-AVS**, a ground-truth-supervised audio-visual benchmark with discrete labels (True/False or 0–5). Concord improves agreement with human judgments over single-judge and naive aggregation baselines, and prioritizing low-agreement instances focuses human review on the most ambiguous cases. We use locally hosted open-source judges and include the binary results for online larger scale models GPT-4.0 mini turbo and Gemini 3.1 Flash Lite.

1 Introduction

Large language models (LLMs) are increasingly deployed in settings where outputs are free-form or grounded in rich context, including multimodal inputs such as audio and video. Yet evaluating these systems remains a central bottleneck. Human evaluation can be expensive, slow to scale, and inconsistent across raters, while reference-based automatic metrics can be unreliable when outputs are free-form or require nuanced matching to a reference. As a result, practitioners increasingly turn to *LLM-as-a-judge* evaluation to obtain scalable assessments and structured feedback.

However, single-judge pipelines are not fully reliable: prior work (Chen et al., 2024; Zheng et al., 2023; Li et al., 2024) has documented systematic judge behaviors (e.g., sensitivity to output order and verbosity) and imperfect alignment with human judgments. A single model's score provides limited visibility into evaluator uncertainty, making it difficult to identify ambiguous instances and allocate additional supervision where it is most needed.

We present **Concord**, a multi-adjudication evaluation pipeline that treats each judge model as an annotator and aggregates their discrete judgments into a single evaluation output. For each evaluation item, judges provide either (i) a binary correctness label or (ii) an ordinal correctness score on a 0–5 scale, depending on the task format. Concord summarizes the panel's responses and uses inter-judge agreement as a practical uncertainty signal to flag potentially ambiguous items for additional review. This enables *disagreement-driven triage*: low-agreement instances can be escalated to human raters (or to a stronger judge), concentrating supervision on the cases most likely to be contentious. In this view, disagreement becomes an actionable indicator of reliability rather than noise to be ignored.

We evaluate Concord on a ground-truth-supervised *multimodal* benchmark (audio-visual inputs (Chen et al., 2020)), where the evaluation signal is discrete (True/False or 0–5), enabling a unified protocol for correctness-oriented assessment. Our experiments measure: (i) correctness against ground truth compared to single-judge baselines and simple aggregation rules, (ii) agreement-with-human metrics (e.g., MCC and Cohen's κ) and ordinal-score alignment (MAE) where applicable, and (iii) the practical utility of disagreement-driven triage under a fixed human-review budget by prioritizing low-agreement instances for review. Concord is model-agnostic: it can be instantiated with locally hosted open-source judges, and it can optionally incorporate a stronger proprietary judge within the same pipeline for escalation or comparison.

We have three main contributions. First, we pro-

* These authors contributed equally.

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 502–510

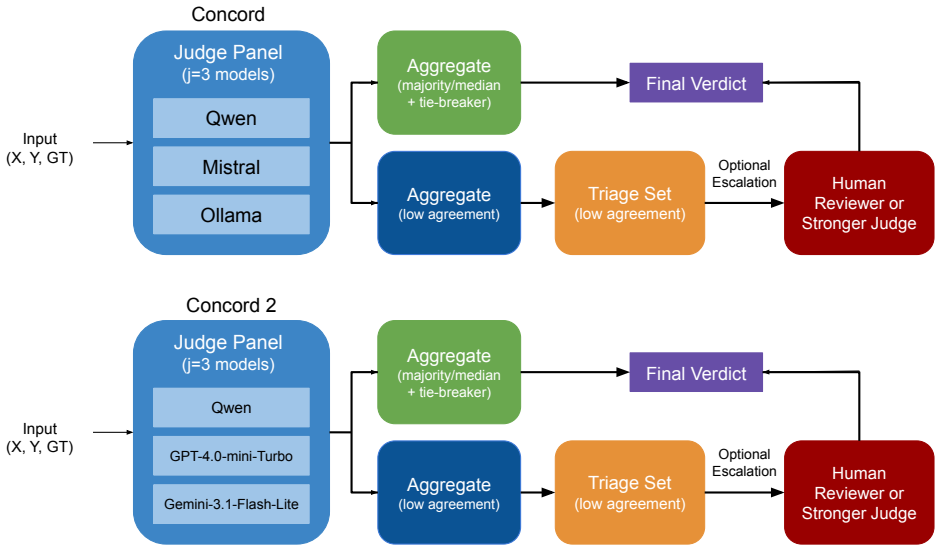

Figure 1: Concord overview. Given an example (X, Y, GT), a panel of J LLM judges produces discrete judgments that are aggregated into a final verdict (majority vote for binary labels or median for 0–5 scores), with Qwen used as a deterministic tie-breaker. In parallel, Concord computes an agreement signal (vote margin or score dispersion) to form a low-agreement triage set, which can be optionally escalated to a human reviewer or stronger judge for arbitration.

pose **Concord**, a multi-adjudication LLM evaluation pipeline that aggregates a local jury's discrete judgments into a single binary/ordinal evaluation output. Next, we empirically evaluate Concord on **AVSSD** (Chen et al., 2020) (human-judged subset, n=174) and report correctness and **SCORE-AVS** (human-judged subset, n=300), agreement-with-human metrics, and ordinal-score alignment. Finally, we examine a **human-in-the-loop** workflow in which low-agreement items are prioritized for review under a fixed budget, illustrating how disagreement can focus human effort on the most ambiguous cases.

2 Related Works

LLM-as-a-Judge and Its Limitations. LLM-as-a-judge has become a common approach for scalable evaluation when outputs are free-form or when exact-match metrics are inadequate. However, recent works document systematic judge artifacts, including sensitivity to answer ordering and verbosity, as well as imperfect alignment with human judgments in some settings (Chen et al., 2024; Zheng et al., 2023; Li et al., 2024). Tools like EvalLLM study how practitioners design and operationalize LLM-based evaluation criteria in human-in-the-loop workflows (Desmond et al., 2024).

Multi-Judge and Multi-Agent Evaluation. To improve robustness beyond a single evaluator, recent methods use multiple LLMs through panels or structured interaction. ScaleEval uses multi-round agent debate to meta-evaluate LLM evaluators under diverse scenarios (Chern et al., 2024). JudgeBench provide systematic testbeds for studying the reliability of LLM-based judges across challenging domains such as reasoning and coding (Tan et al., 2024). These directions motivate treating judges as a population of annotators rather than relying on a single model's decision.

Uncertainty signals and selective evaluation. A separate line of work studies when to *trust* a cheaper judge versus *escalate* to stronger supervision, often emphasizing selective evaluation and escalation policies (Jung et al., 2024). Other approaches align judge scoring criteria with human labels via iterative refinement (Liu et al., 2024) or mixed-initiative systems for building evaluation functions with human feedback (Shankar et al., 2024). In contrast, Concord focuses on ground-truth-supervised settings with discrete evaluation signals and uses inter-judge agreement as a practical uncertainty signal to drive triage, prioritizing low-agreement instances for additional review.

3 Method: Concord

Concord is an ensemble-based evaluation pipeline for *correctness-oriented* tasks with available ground truth and *discrete* evaluation signals. There

were two different implementations of Concord, one using local offline models. For the first implementation, we focus on settings where the evaluator output is either a binary correctness label (True/False) or an ordinal score on a bounded scale (e.g., 0–5), while the second implementation used only the binary correctness label. Although inputs may be multimodal and candidate answers may be free-form text, the evaluation target is discrete and grounded in a reference signal. Concord is not designed for preference-based evaluation of long-form reasoning or descriptive explanations; instead, it provides (i) an aggregated verdict and (ii) an agreement-based uncertainty signal for selective review. Figure 1 illustrates how Concord aggregates judge outputs and performs agreement-based triage.

Setup and Judge Panel. For each example we consider (X, Y, GT), where X denotes the task input (potentially multimodal), Y is the candidate model output to be evaluated, and GT is the ground-truth reference used for correctness judging. Concord queries a panel of J judge models using a shared rubric and a standardized prompt template. Each judge outputs a discrete judgment: either a binary label $y_j \in \{0, 1\}$ or an ordinal score $s_j \in \{0, 1, 2, 3, 4, 5\}$, depending on the benchmark.

Ensemble Aggregation with Tie-Breaking. Concord aggregates the panel outputs into a single verdict. Rather than relying on a simple majority vote, Concord considers all judge outputs jointly to produce a final decision, while remaining grounded in the sub-judge predictions. In cases of full agreement, Concord follows the unanimous decision. When the panel is maximally uncertain (i.e., tied), we apply a deterministic tie-breaking rule that gives priority to Qwen. For binary evaluation, let $c_1 = \sum_{j=1}^{J} \mathbb{I}[y_j = 1]$ and $c_0 = J - c_1$. The aggregated verdict is

$$\hat{y} = \begin{cases} 1 & \text{if } c_1 > c_0, \\ 0 & \text{if } c_0 > c_1, \\ y_{\text{Qwen}} & \text{if } c_1 = c_0. \end{cases}$$

For ordinal evaluation, we set $\hat{s} = \text{median}(\{s_j\}_{j=1}^{J})$ and break non-unique medians (e.g., even J) using s_{Qwen}. We choose Qwen for tie-breaking as it has the strongest standalone agreement with human labels among our local judges (§4.2), and using it only on ties preserves the ensemble while ensuring deterministic outputs.

Agreement as an uncertainty signal and selective triage. Beyond the final verdict, Concord computes an agreement signal across judges that serves as a proxy for uncertainty. This signal is used to identify borderline examples for selective evaluation (e.g., escalation to human review or an additional judge). Triage is computed from the panel distribution, providing a model-agnostic view of uncertainty.

In the binary setting, we use majority voting and flag examples for review when there is disagreement among judges (e.g., a 2–1 split with $J = 3$). In the ordinal setting, we similarly flag examples with high variation in scores, indicating uncertainty in the assigned rating.

Disagreement-based human triage. In addition to agreement-based signals, we introduce a complementary rule to capture cases where the panel disagrees with the final decision. Concord uses Qwen as the final adjudicator informed by all judges, but we flag examples where at least two judges disagree with the final prediction.

This identifies low-confidence or ambiguous cases where the final decision may be unreliable. The union of these criteria forms the "disagreement" subset used in our human-in-the-loop experiments. In practice, this approach concentrates human evaluation on contentious examples, reducing annotation cost while improving reliability.

Implementation and outputs. We run each judge model independently and aggregate their outputs offline: per-judge predictions are exported to CSV, then combined to produce the final verdict ($\hat{y}$ or $\hat{s}$) and a triage list of flagged low-confidence examples. We then compare against human labels and report standard correctness and agreement metrics. The codebase is available at `https://github.com/BASHLab/concord`.

4 Evaluation

4.1 Experimental Setup

Benchmark and Protocol. We evaluate Concord on AVSSD and SCORE-AVS using a correctness-oriented protocol with ground truth and discrete supervision. Both datasets inputs may include audio-visual context and candidate answers may be free-form, each instance is evaluated as either (i) a binary correctness label (True/False) or (ii) an ordinal score on a 0–5 scale, depending on the task variant. Concord is instantiated with a locally hosted panel of open-source judges and outputs an aggregated verdict for each example.

Table 1: **AVSSD** results on the human-judged subset ($n = 174$), consisting of 100 standard examples and 74 disagreement-heavy examples where at least two judges disagree. Binary judging metrics are Accuracy (Acc), Precision (Prec), and Recall (Rec). Agreement with human labels is measured by MCC and Cohen's κ. Ordinal scoring reports MAE and mean score. **Best values per column are bolded** (MAE is minimized; Mean is closest to the human mean). While the first 100 examples reflect typical performance where strong individual judges perform well, the inclusion of disagreement-heavy cases highlights robustness: Concord maintains high performance across the full dataset, demonstrating improved reliability in ambiguous settings where single judges degrade.

Method	**Binary**			**Agreement**		**Ordinal**	
	Acc	Prec	Rec	MCC	κ	MAE	Mean
Concord	**93.10**	94.80	**96.20**	**0.806**	**0.805**	**0.580**	**3.52**
B1: Best single judge (Qwen)	82.80	97.20	79.70	0.635	0.602	0.621	3.25
Mistral	54.00	**98.20**	40.60	0.348	0.230	0.937	2.94
Ollama	59.80	87.10	55.60	0.245	0.205	0.989	3.08

Table 2: **Score_Avs** results on the full combined dataset ($n = 300$), a subset of the first 100 samples, and 200 samples with 2 model disagreement, after dropping NaN values. This dataset includes all subsets (standard and disagreement-heavy). Binary judging metrics are Accuracy (Acc), Precision (Prec), and Recall (Rec). Agreement with human labels is measured by MCC and Cohen's κ. Ordinal scoring reports MAE and mean score. **Best values per column are bolded** (MAE is minimized; Mean is closest to the human mean of 0.73).

Method	**Binary**			**Agreement**		**Ordinal**	
	Acc	Prec	Rec	MCC	κ	MAE	Mean
Concord	**74.30**	82.60	**82.20**	**0.351**	**0.351**	1.487	**0.727**
B1: Best single judge (Qwen)	51.70	93.00	36.50	0.286	0.192	1.093	0.287
Mistral	46.30	**96.80**	27.40	0.273	0.155	**0.950**	0.207
Qwen	51.70	93.00	36.50	0.286	0.192	1.093	0.287
Ollama	63.00	76.70	70.80	0.121	0.120	1.430	0.673

Baselines and Metrics. We compare Concord against a single baseline: **B1**, a single local judge instantiated as a 7B Qwen model. Due to lack of access to proprietary models, we do not include a proprietary judge baseline. We report correctness metrics (accuracy, precision, recall, F1), agreement with human labels (MCC and Cohen's κ), and ordinal alignment (MAE).

Selective Triage. Concord uses inter-judge agreement as an uncertainty signal to prioritize ambiguous cases for review. We flag examples with low consensus using the panel's response distribution (e.g., small vote margin for binary labels or high score dispersion for 0–5 ratings), forming a low-agreement subset for the human-in-the-loop analysis.

Implementation. To support reproducibility and allow fully local evaluation, we host the entire Concord jury on local GPU infrastructure. Experiments were run on the Turing HPC cluster and the Raven GPU cluster using NVIDIA A100 and H100 nodes.

4.2 Results

Correctness as a judge (binary). We first evaluate Concord as a correctness judge on the AVSSD human-labeled subset ($n = 174$), measuring whether each response is classified as correct or incorrect relative to ground truth. Table 1 reports standard correctness metrics for Concord and individual judges. Overall, Concord achieves the strongest performance across accuracy (93.10), recall (96.20), and agreement metrics, while maintaining high precision comparable to strong individual judges. This supports the motivation that aggregation mitigates idiosyncratic judge errors without relying on a single evaluator. Per-judge dashboards are provided in Appendix A.

Performance on random subset. We randomly chose 100 samples and have annotators manually label them. Concord matches the strongest individual judge (Qwen) across all binary metrics, achieving identical performance (Accuracy 0.940, MCC 0.875, κ 0.872). This indicates that in relatively unambiguous settings, Concord preserves the strengths of the best individual judge without degradation. Notably, weaker judges such as Ollama and

Table 3: **AVSSD (Disagreement Subset)** results on disagreement-heavy examples ($n = 74$), where at least two judges disagree. Binary judging metrics are Accuracy (Acc), Precision (Prec), and Recall (Rec). Agreement with human labels is measured by MCC and Cohen's κ. Ordinal scoring reports MAE and mean score. **Best values per column are bolded** (MAE is minimized; Mean is closest to the human mean of 0.919).

Method	Binary			Agreement		Ordinal	
	Acc	Prec	Rec	MCC	κ	MAE	Mean
Concord	**91.90**	91.90	**100.00**	0.000	0.000	**0.689**	**1.000**
B1: Best single judge (Qwen)	67.60	**95.80**	67.60	**0.196**	**0.136**	0.824	0.649
Mistral	8.10	0.00	0.00	0.000	0.000	1.365	0.000
Qwen	67.60	**95.80**	67.60	**0.196**	**0.136**	0.824	0.649
Ollama	29.70	94.40	25.00	0.053	0.017	1.351	0.243

Table 4: **Score_Avs (Disagreement Subsets)** results on disagreement-heavy examples ($n = 200$), constructed from two subsets where at least two judges disagree. Binary judging metrics are Accuracy (Acc), Precision (Prec), and Recall (Rec). Agreement with human labels is measured by MCC and Cohen's κ. Ordinal scoring reports MAE and mean score. **Best values per column are bolded** (MAE is minimized; Mean is closest to the human mean of 0.79).

Method	Binary			Agreement		Ordinal	
	Acc	Prec	Rec	MCC	κ	MAE	Mean
Concord	**67.00**	78.00	**81.00**	-0.050	-0.050	1.800	**0.820**
B1: Best single judge (Qwen)	39.00	87.50	26.60	**0.117**	**0.063**	0.900	0.240
Mistral	29.00	**90.00**	11.40	0.090	0.030	**0.880**	0.100
Qwen	39.00	87.50	26.60	**0.117**	**0.063**	0.900	0.240
Ollama	51.00	72.10	62.00	-0.248	-0.239	1.480	0.680

Mistral exhibit lower agreement and higher error rates, highlighting the benefit of combining multiple evaluators.

Performance on disagreement subset. We next evaluate performance on a subset of examples characterized by higher disagreement (74 samples). These samples were selected based on intra-model agreement. With 2 or more models in disagreement with the final adjudicator, each row is appended for human evaluation. In this setting as table 3 shows, Concord significantly outperforms all individual judges, achieving 0.919 accuracy compared to 0.676 for Qwen and substantially lower scores for other models. Concord also achieves perfect recall (1.000), indicating that it successfully captures all positive cases even under high ambiguity. In contrast, individual judges exhibit degraded performance, including near-random or highly biased predictions (e.g., Mistral predicting all negatives). These results demonstrate that Concord is particularly robust in challenging, disagreement-heavy scenarios where single judges are unreliable.

On the two-model disagreement subset of the $Score_Avs$ shown in Table 4, we can see that Concord still outperforms all other individual models. It has the best scores in accuracy (67.00), recall (81.00) and mean (0.820). Although Mistral and Qwen outperform Concord in metrics such as precision, MCC, Cohen's kappa, and MAE, Concord demonstrates overall consistent performance.

Agreement with human annotations. Beyond correctness, we measure statistical reliability using MCC and Cohen's κ (Table 1). On the full dataset, Concord achieves the highest agreement with human labels (MCC 0.806, κ 0.805), substantially exceeding weaker judges whose outputs are less correlated with human annotation. While agreement metrics degrade on the disagreement subset due to class imbalance, Concord maintains strong correctness performance, indicating robustness even when standard agreement measures become less informative.

Ordinal scoring (0–5). We additionally evaluate ordinal correctness scores on a 0–5 scale. As shown in Table 1, Concord achieves low MAE (0.580) relative to human scores and closely matches the mean of the human rating distribution (3.52 vs. 3.45), suggesting that aggregation counteracts systematic over- or under-scoring by individual judges.

Generalization to score_avs (n=300). To further evaluate robustness, we introduce a separate dataset, *score_avs*, constructed from a larger pool of approximately 3,000 examples. From this pool, we include (i) the first 100 standard examples and

Table 5: **Human-in-the-loop triage on disagreement-heavy subsets.** We evaluate selective review on AVSSD ($n = 74$) and Score_Avs ($n = 100$) disagreement subsets. ECR (Error Correction Rate) measures the number of model errors identified per 100 human reviews. Concord prioritizes disagreement cases, while Random samples uniformly. *Random baseline values are estimated from overall dataset error rates rather than computed from explicit sampled subsets.*

Dataset	Strategy	Budget B	# errors found	ECR ↑
AVSSD (disagreement)	Concord triage	74	6	**8.1**
AVSSD (disagreement)	Random (est.)	74	∼5	6.9
Score_Avs (disagreement)	Concord triage	100	33	**33.0**
Score_Avs (disagreement)	Random (est.)	100	∼26	26.0

(ii) 200 additional examples specifically selected where at least two models disagree, yielding a total of $n = 300$. This dataset therefore contains a substantially higher proportion of ambiguous and disagreement-heavy cases.

As shown in Table 2, Concord continues to outperform all individual judges on this more challenging dataset, achieving the highest accuracy (0.743), recall (0.822), and agreement metrics (MCC/κ = 0.351), while maintaining close alignment with the human label distribution (mean 0.727 vs. 0.730). In contrast, individual judges degrade substantially: Mistral and Qwen exhibit low recall (0.274 and 0.365), while Ollama, despite higher recall, shows weaker agreement (MCC 0.121). These results confirm that Concord remains effective even when the evaluation distribution is explicitly biased toward difficult, disagreement-heavy examples.

Why not just use Qwen as the judge? Qwen is the strongest single local judge on standard examples, but its performance degrades substantially on disagreement-heavy subsets (e.g., accuracy drops from 0.940 to 0.676). In contrast, Concord maintains high performance (0.919 accuracy) under the same conditions and continues to outperform on the *score_avs* dataset (0.743 vs. 0.517 accuracy). Thus, Concord is preferable when robustness and generalization are required, particularly in ambiguous or high-variance settings.

Human-in-the-Loop triage. Finally, we test whether agreement-based triage can focus limited human effort on the most ambiguous cases. We use a fixed review budget of $B = 100$ and compare Concord triage, which selects high-disagreement examples. We measure efficiency using the Error Correction Rate (ECR), defined as the number of model errors corrected per 100 human reviews. Table 5 demonstrates that with a randomly selected sample size, on AVSSD only about 5 errors were calculated for, while using the Concord triage with intra-model disagreement, 6 were accounted for resulting in more instances caught for humans to evaluate. Similarly, a randomly sampled subset from Score_AVS accounted for 26 errors while the Concord triage accounted for 33.

When applied to large-scale datasets (N>3,000), this reduces the review space by approximately 98% (reviewing 100 items instead of the full set). Agreement- and disagreement-based triage concentrates review near the decision boundary, prioritizing cases where judges disagree most.

Online models in judge panel. We also evaluate Concord with stronger online models by substituting Mistral and Ollama with GPT-4.0-mini-Turbo and Gemini-3.5-Flash-Lite. In this setting, the judges agree on the majority of examples, with only 184 disagreements out of 3,000 total samples. This indicates that higher-quality models lead to more consistent judgments, correctly resolving most cases without additional aggregation.

However, these remaining disagreement cases are precisely where Concord provides value. Rather than relying on a single model, Concord identifies and focuses on these ambiguous instances, where model predictions diverge. In this subset, Concord continues to provide meaningful improvements by resolving a substantial portion of disagreements while also triggering human-in-the-loop triage for the most uncertain cases.

These results suggest that as individual models improve, the role of Concord shifts from general evaluation toward targeted adjudication. Even with strong base models, disagreement persists in a non-trivial number of cases, and Concord provides a principled mechanism to resolve or escalate these instances, improving overall reliability.

5 Conclusion, and Future Work

Concord is an ensemble-based, multi-adjudication evaluation pipeline for correctness-oriented tasks

with discrete ground truth signals. By using inter-judge agreement as a practical uncertainty signal, Concord supports disagreement-driven triage: low-consensus cases can be escalated for human review, concentrating supervision where automated judging is least reliable. Our results show that a locally hosted panel of open-source judges can achieve strong agreement with human labels (e.g., MCC 0.806, κ 0.805) while outperforming individual judges in overall reliability and calibration. In particular, Concord improves recall and ordinal alignment while maintaining high precision, demonstrating that aggregation mitigates individual judge biases. This enables meaningful reductions in manual review under a fixed budget by prioritizing ambiguous cases.

Future work includes exploring weighted aggregation and judge-specific reliability modeling, improving efficiency when incorporating richer judge rationales. In future, we plan to evaluate the performance of utilizing a wide variety online judges including GPT-5.

6 Limitations

Our work is subject to several practical limitations. First, the majority of experiments rely on locally hosted, open-source models available through free Hugging Face transformers. While this enables reproducibility and low-cost experimentation, it restricts the strength and diversity of judges compared to larger proprietary models.

Second, experiments with online models (e.g., GPT-4.0-mini-Turbo and Gemini-3.5-Flash-Lite) were constrained by both cost and runtime. API usage introduces latency and financial overhead, limiting the scale of evaluation and the number of configurations that can be explored. In practice, this makes large-scale ensemble evaluation with online models less feasible without significant resources.

Third, the dataset size is relatively limited. Although we construct targeted disagreement subsets to stress-test robustness, the total labeled data remains small compared to real-world deployment settings. Additionally, generating judge outputs across multiple models is computationally expensive, further constraining dataset expansion.

Finally, experiments were conducted on limited GPU resources, which restricted batch sizes, model selection, and overall throughput. Running multiple judges in parallel is inherently resource-intensive, and scaling Concord to larger datasets or stronger models would require more substantial compute infrastructure.

These limitations highlight trade-offs between cost, scalability, and model quality, and motivate future work on more efficient aggregation strategies and scalable evaluation pipelines.

7 Ethics Statement

Concord improves evaluation reliability through aggregation, but it inherits biases present in its underlying models. While combining multiple judges reduces variance, systematic errors shared across models may persist, particularly in ambiguous or subjective cases.

Automated evaluation should not replace human judgment in high-stakes settings. Concord is designed to support human oversight, with disagreement-based triage explicitly identifying uncertain cases for review.

Finally, the framework relies on limited labeled data and computationally intensive models, which may introduce bias, cost, and accessibility concerns. Concord should therefore be deployed with appropriate safeguards and human-in-the-loop validation.

References

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *arXiv preprint arXiv:2401.16788*.

Michael Desmond, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M Johnson. 2024. Evalullm: Llm assisted evaluation of generative outputs. In *Companion proceedings of the 29th international conference on intelligent user interfaces*, pages 30–32.

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2024. Trust or escalate: Llm judges with provable guarantees for human agreement. *arXiv preprint arXiv:2407.18370*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Calibrating llm-based evaluator. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pages 2638–2656.

Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Per-Judge Dashboards and Diagnostics on AVSSD

A.1 Dashboards on the human-labeled subset ($N = 174$)

To complement aggregate metrics, we provide compact, model-specific dashboards summarizing evaluator behavior on the AVSSD human-labeled subset ($N = 174$), which includes 100 standard examples and 74 disagreement-heavy examples.

Each dashboard reports: (i) binary correctness metrics (accuracy, precision, recall, F1), (ii) a confusion matrix, and (iii) ordinal semantic scoring on a 0–5 scale (model mean, human mean, and MAE). Figures 2–5 visualize these summaries for Concord and individual judges.

A.2 Subset analysis (standard vs. disagreement)

We analyze performance separately on the standard subset (first 100 examples) and the disagreement-heavy subset (74 examples where at least two judges disagree). On the standard subset, Concord matches the strongest individual judge (Qwen), achieving identical performance across binary metrics (Accuracy 0.940, MCC 0.875).

On the disagreement subset, individual judges degrade substantially, while Concord remains stable. For example, Qwen's accuracy drops to 0.676, whereas Concord maintains 0.919 accuracy with perfect recall (1.000). Other judges exhibit more extreme behavior, such as Mistral predicting predominantly a single class. The disagreement subset was obtained by adding the three predicted binary values and checking if it was not equal to three or zero.

A.3 Summary of judge behavior

The dashboards highlight complementary behaviors across judges. Qwen performs strongly on standard examples, Mistral tends to be conservative, and Ollama exhibits higher variability. By aggregating these signals, Concord reduces individual biases and maintains consistent performance across both standard and disagreement-heavy subsets.

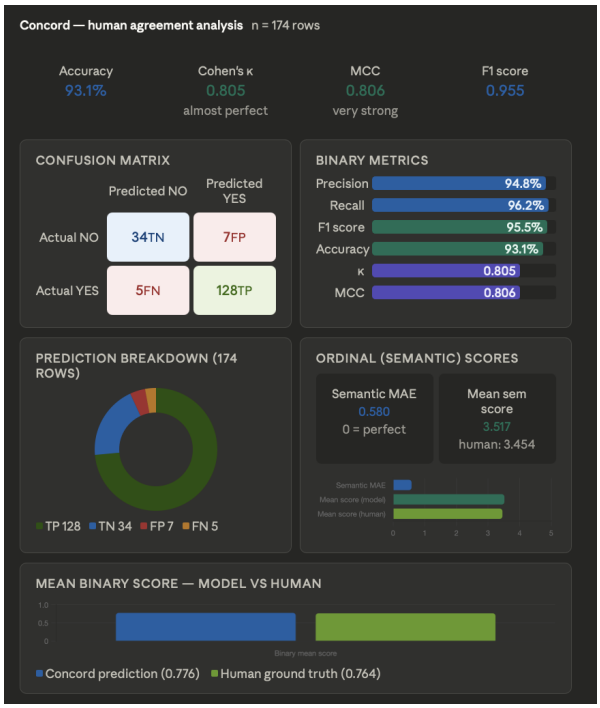

Figure 2: **Concord dashboard on AVSSD** ($N = 174$)**.** Summary of binary correctness, confusion matrix, and ordinal scoring, showing strong alignment with human labels across both standard and disagreement-heavy cases.

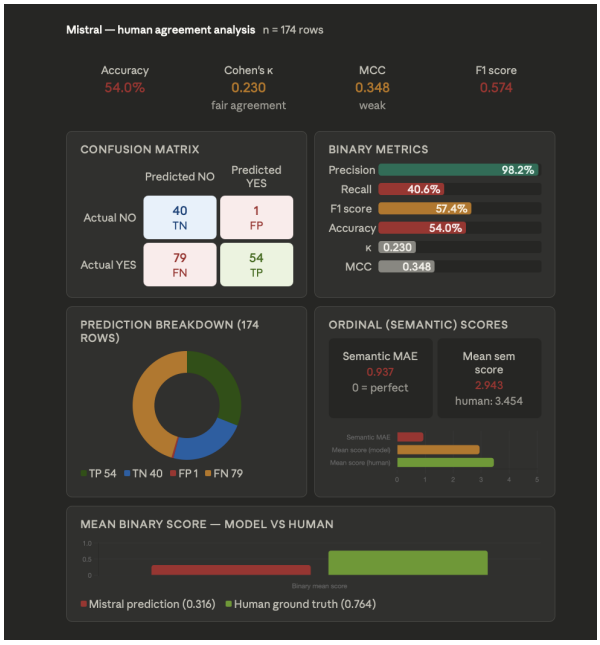

Figure 3: **Mistral dashboard on AVSSD** ($N = 174$)**.** Metrics and confusion matrix illustrating conservative predictions and reduced recall, especially on disagreement-heavy examples.

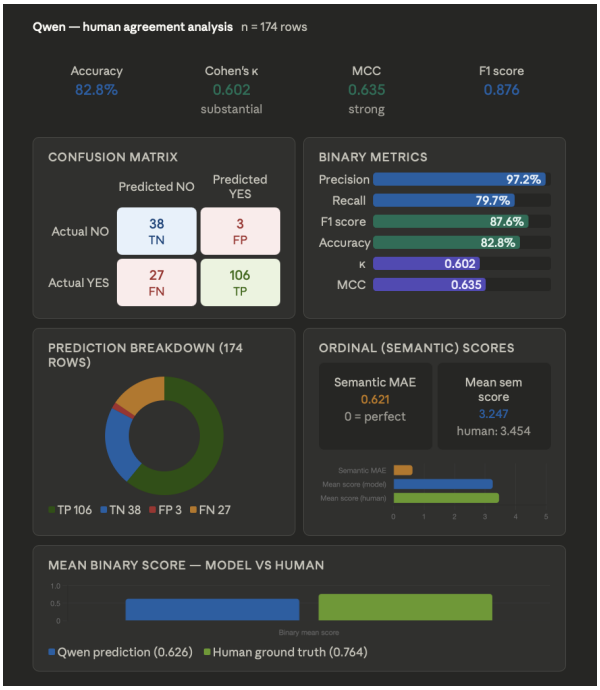

Figure 4: **Qwen dashboard on AVSSD** ($N = 174$)**.** Strong performance on standard examples with reduced reliability under disagreement.

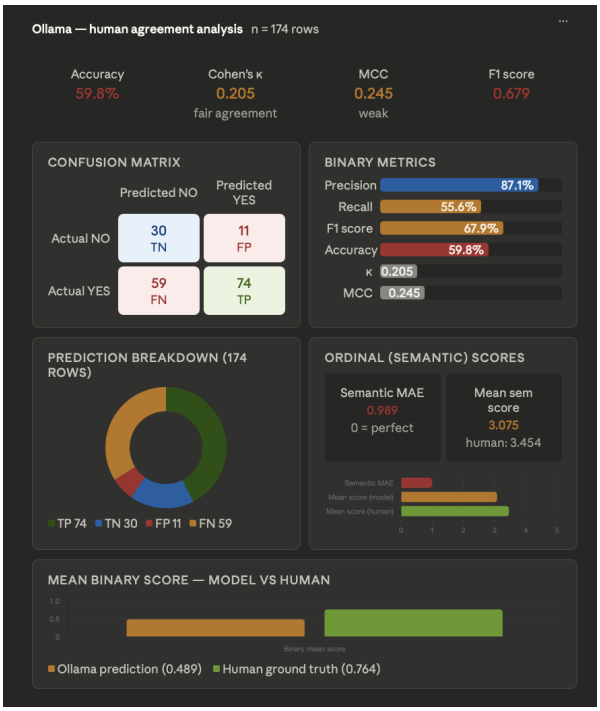

Figure 5: **Ollama dashboard on AVSSD** ($N = 174$)**.** Higher recall but increased variability and false positives on disagreement-heavy cases.