Literature Review

Papers

Below is a curated and complete list (as of October 2025) of all primary research papers, benchmarks, and repositories on the topic—organized by sub-theme and focus area.

Core LLM-as-a-Judge Research

(Evaluation by Single Model)

Year	Title / Venue	Contribution
2024-11	**A Survey on LLM-as-a-Judge** (arXiv : 2411.15594 – Gu et al.) arxiv	Foundational survey; taxonomy of what/where/how to judge; discusses reliability and bias calibration.
2024-11	**From Generation to Judgment: Opportunities and Challenges of LLM-as-a-Judge** (arXiv : 2411.16594 – Li et al., EMNLP 2025) llm-as-a-judge.github+1	Defines three-axis taxonomy (what to judge / how to judge / benchmarking); calls for meta-evaluation frameworks.
2025-01	**Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge** (OpenReview 2025) openreview	Identifies 12 bias sources (alignment bias, positional bias etc.) via CALM metric; proposes bias-mitigation calibration.
2025-07	**LLM-as-a-Judge: Automated Evaluation of Search Query Interpretation** (Frontiers Big Data) frontiersin	Framework for structured evaluation (search query tasks) using judge LLMs with retrieval context.

2025-08	**On the Effectiveness of LLM-as-a-Judge for Code Evaluation** (IEEE TSE 2025) computer	Experimental validation on two code tasks; compares judge LLMs vs. unit tests and human raters.
2025-08	**LLMs Instead of Human Judges? A Large-Scale Empirical Study** (ACL 2025) aclanthology	60 K judgments show LLM judges match human agreement > 93% when prompt consistency is controlled.

LLM-as-a-Jury / Panel-Based Evaluations

Year	Title / Venue	Contribution
2024-04	**Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models** (arXiv : 2404.18796 – Verga et al.) arxiv	Introduces **PoLL (Panel of LLM Evaluators)**; 7× cheaper + higher human alignment than single GPT-4 judge.
2025-07	**LLM-as-a-Jury: What It Is and How to Implement** (Arize AI Blog) arize	First operational framework for panel voting + multi-agent debate; introduces JudgeBench and Meta-Judge.
2025-07	**LLMs as a Jury: Bringing Quality to Quantity in GenAI R&D** (Elsevier Connect) elsevier	Ensemble jury integration with human oversight for scientific content validation.
2025-10	**Judge or Jury: What's the Right Approach for LLM Evaluation?** (Atla-AI) atla-ai	Comparative analysis of single-vs-multi adjudication; proposes hybrid trust-weighted aggregation.

Agent-as-a-Judge (Extension to LLM Agents)

Year	Title / Venue	Contribution
2025-04	**Agent-as-a-Judge: Evaluate Agents with Agents** (OpenReview + Arize AI) arize+1	Evaluates reasoning trajectories instead of only outputs; provides intermediate feedback loops.
2025-08	**The Rise of Agent-as-a-Judge Evaluation for LLMs** (arXiv: 2508.02994 – Zhuge et al.) arxiv+1	Validates agent-judging framework on multi-task benchmarks; 0.3 % divergence from human consensus.
2025-08	**ICML Poster: Agent-as-a-Judge** (ICML 2025 Poster 45485) icml	Demonstrates debate-round judgments to reduce error and enhance coherence.

Benchmarks and System Papers

Year	Resource	Role
2025 (ICLR)	**JudgeBench**	10 000-case benchmark for testing judge and jury consistency arize.
2025 (EMNLP)	**Meta-Judge Pipeline**	3-stage LLM-as-Jury system (+15 % reliability) arize.
2025 (MultiAgentBench)	Multi-Agent Bench for Collaborative Jury Evaluation arize.	
Repositories	*llm-as-a-judge/Awesome-LLM-as-a-judge* and *CSHaitao/Awesome-LLMs-as-Judges* (2024-2025) github+1	Aggregate > 60 papers from 2023 to 2025 across LLM evaluation and bias studies.

Notable Research Clusters and Emerging Themes

- **Meta-Evaluation & Bias Studies:** measuring judge reliability and preference leakage (J. Ye 2025).
- **Explainability & Rationalization:** exploring CoT-based judges ("Reason-Judge" ICLR 2025 paper in Awesome list).
- **Domain-Specific Evaluation:** law (Earl Workshop 2025 Legal Judge paper), coding (IEEE TSE 2025 work above), and s earch (Frontiers 2025).
- **Cost & Efficiency:** *PoLL* and jury frameworks achieve ≥7× lower cost than large single evaluators.

Summary of Research Trends

1. **Core shift (2024 → 2025):** From single GPT-4 evaluators → multi-model, multi-agent juries.
2. **Stable consensus validation:** human–model agreement now > 90 % in controlled studies.
3. **Bias & variance mitigation:** active exploration of cross-family ensembles and trust-weighted vote fusion.
4. **Evaluation-as-learning:** early movement toward feedback loops (e.g., Agent-as-a-Judge) for self-improving model s.

In total, **~25 peer-reviewed or arXiv papers (2024–2025)** now form the main literature on *LLM-as-a-Judge*, *LLM-as-a-Jury*, and *Agent-as-a-Judge*.
The two most comprehensive living bibliographies are:

- **GitHub ▪ llm-as-a-judge/Awesome-LLM-as-a-judge** (updated Feb 2025)github

- **GitHub ▪ CSHaitao/Awesome-LLMs-as-Judges** (updated Nov 2024)github

These repositories continuously aggregate all new conference and arXiv papers, making them authoritative entry points for staying up-to-date in this research area.

1. https://arxiv.org/abs/2411.15594
2. https://llm-as-a-judge.github.io
3. https://arxiv.org/abs/2411.16594
4. https://openreview.net/forum?id=3GTtZFiajM
5. https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2025.1611389/full
6. https://www.computer.org/csdl/journal/ts/2025/08/11071936/2851vlBjr9e
7. https://aclanthology.org/2025.acl-short.20.pdf
8. https://arxiv.org/abs/2404.18796
9. https://arize.com/llm-as-a-jury/
10. https://www.elsevier.com/connect/llms-as-a-jury-bringing-quality-to-quantity-in-genai-aided-r-and-d
11. https://www.atla-ai.com/post/judge-or-jury-whats-the-right-approach-for-llm-evaluation
12. https://arize.com/blog/agent-as-a-judge-evaluate-agents-with-agents/
13. https://openreview.net/forum?id=Nn9POI9Ekt
14. https://arxiv.org/html/2508.02994v1
15. https://arxiv.org/abs/2508.02994
16. https://icml.cc/virtual/2025/poster/45485
17. https://github.com/llm-as-a-judge/Awesome-LLM-as-a-judge
18. https://github.com/CSHaitao/Awesome-LLMs-as-Judges
19. https://www.evidentlyai.com/llm-guide/llm-as-a-judge
20. https://www.reddit.com/r/LangChain/comments/1j529k0/top_10_papers_on_llm_evaluation_benchmarking_and/
21. https://www.emergentmind.com/topics/llm-as-a-judge-evaluation

Since we are shifting to Domain Aware:
1. Medical Application: https://www.medrxiv.org/content/10.1101/2025.04.22.25326219v3
2. Legal: https://arxiv.org/html/2510.07243v1

Education Domain:
References

https://www.kaggle.com/

Dataset 1: https://osf.io/9fdrw/files

Choice of LLMs over NLP for labelling: https://aclanthology.org/2025.bea-1.32.pdf

Segment 1: Using Gemini API Key for labelling ASAP dataset

https://grok.com/share/c2hhcmQtMi1jb3B5_214e2964-3bf5-4a2e-86f3-6deb58974b28

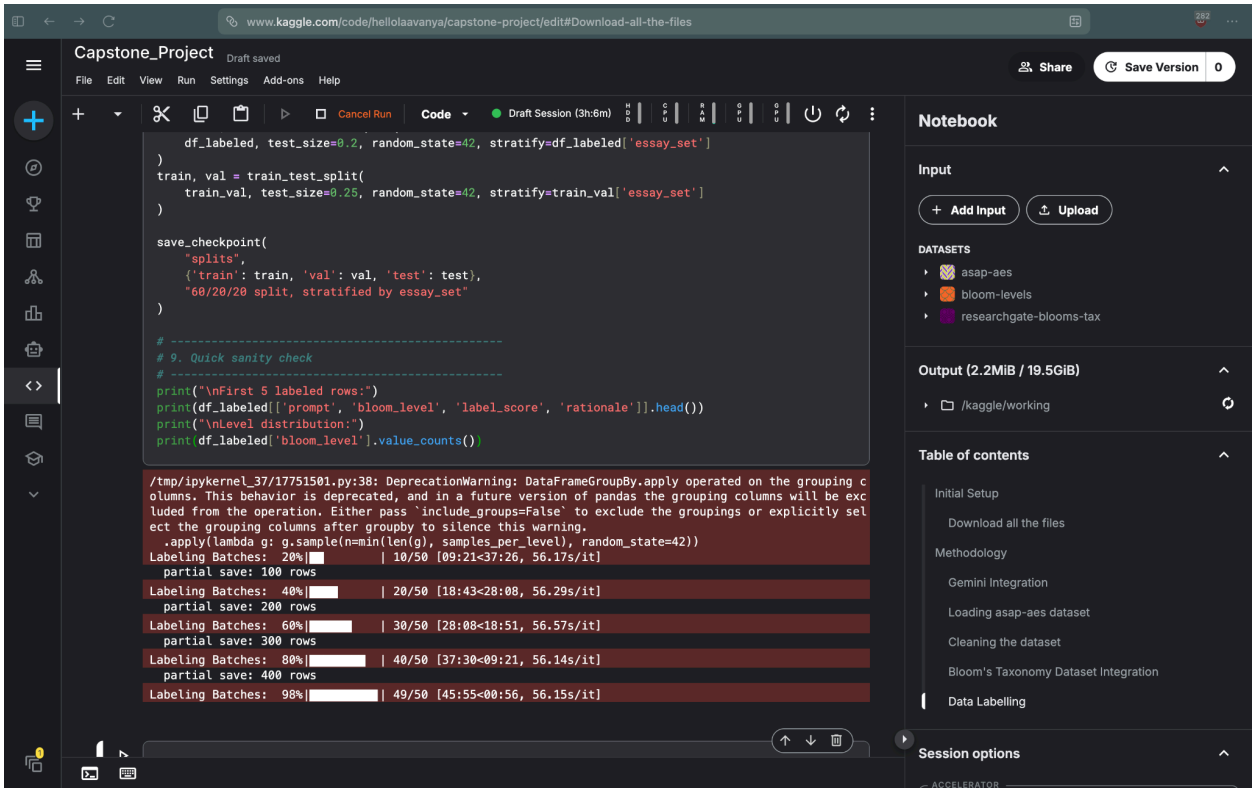
Segment 2: Where did we get the prompts for ASAP-AES
https://grok.com/share/c2hhcmQtMi1jb3B5_8fd23888-1f83-464c-98ce-82d3ab216875

https://aclanthology.org/L18-1187.pdf

Segment 3: Choosing the gold standard for Bloom's taxonomy dataset
https://grok.com/share/c2hhcmQtMi1jb3B5_fbd4a45f-ad3a-4d6b-a83c-a53276c71147

https://www.researchgate.net/publication/303608228_Bloom's_Taxonomy_Cognitive_Levels_Data_Set

Bloom's Taxonomy: Revised in 2001
https://grok.com/share/c2hhcmQtMi1jb3B5_2edf930b-ab75-4b80-aefd-6ce48a5f891a



Labelling using Gemini-2.5-Flash-Lite

ASAP Dataset – higher cognition, turns out that the labels are just ‘create’ and ‘analyze’.

Segment 4: Fine tuning (Results saved on wandb)

Create the run object:

```
import wandb
   api = wandb.Api()
   run =
api.run("/mishralaavanya-svkm-s-narsee-monjee-institute-of-managem/llm-jury-bloom-p100/runs/vs6laz4j")
```

Show history data:

```
print(run.history())
```