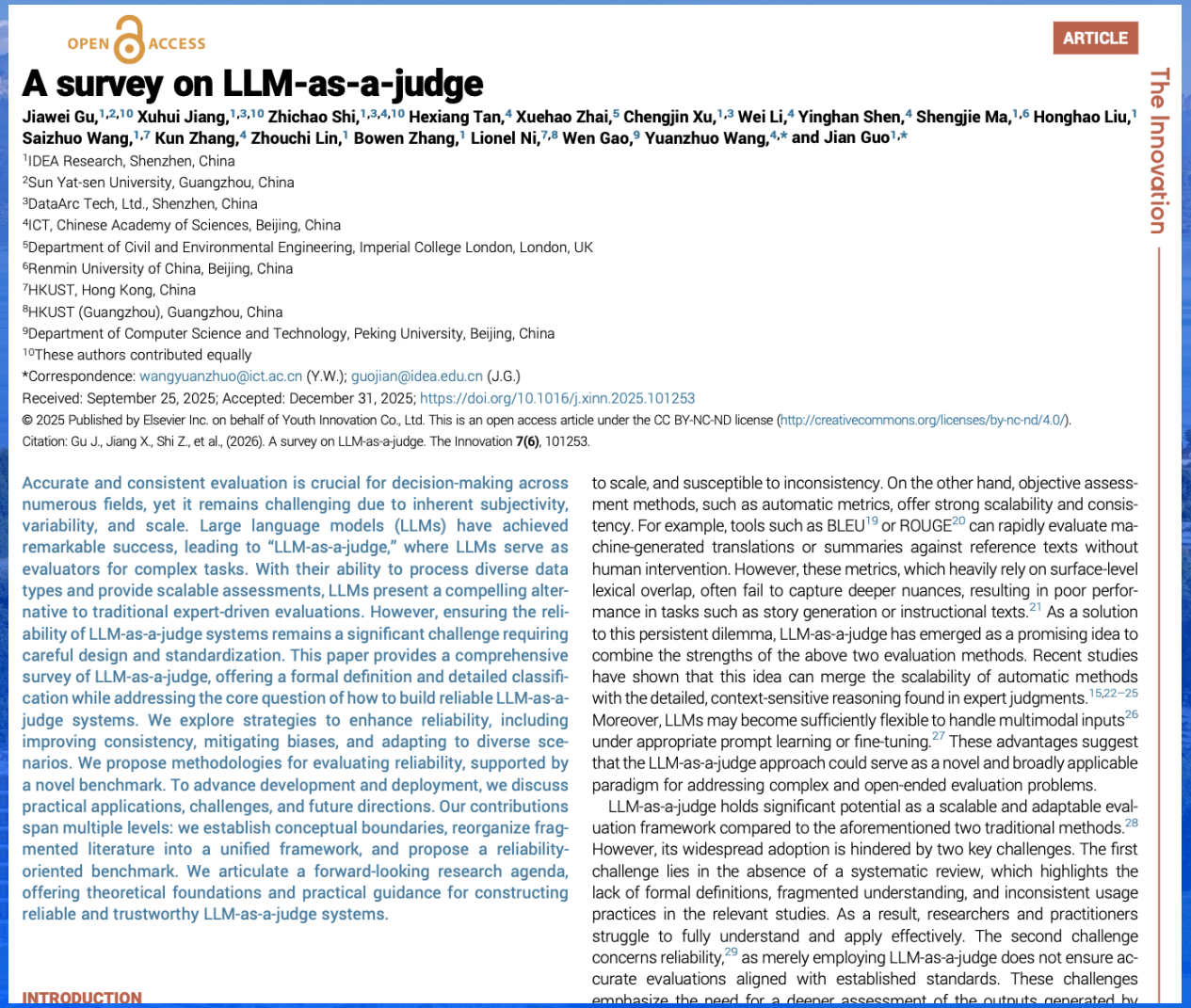
Unil.

Critical Reading Exam

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu H., Wang, S., Zhang, K., Lin, Z., Zhang, B., Ni, L., Gao, W., Wang, Y., & Guo, J. [3]. A Survey on LLM-as-a-Judge. The Innovation, 101253. DOI: 10.1016/j.xinn.2025.101253, 2026

03.06.2026 | Stergios Konstantinidis



OPEN ACCESS

ARTICLE

The Innovation

A survey on LLM-as-a-judge

Jiawei Gu,[1,2,10] Xuhui Jiang,[1,3,10] Zhichao Shi,[1,3,4,10] Hexiang Tan,[4] Xuehao Zhai,[5] Chengjin Xu,[1,3] Wei Li,[4] Yinghan Shen,[4] Shengjie Ma,[1,6] Honghao Liu,[1] Saizhuo Wang,[1,7] Kun Zhang,[4] Zhouchi Lin,[1] Bowen Zhang,[1] Lionel Ni,[7,8] Wen Gao,[9] Yuanzhuo Wang,[4,*] and Jian Guo[1,*]

[1]IDEA Research, Shenzhen, China
[2]Sun Yat-sen University, Guangzhou, China
[3]DataArc Tech, Ltd., Shenzhen, China
[4]ICT, Chinese Academy of Sciences, Beijing, China
[5]Department of Civil and Environmental Engineering, Imperial College London, London, UK
[6]Renmin University of China, Beijing, China
[7]HKUST, Hong Kong, China
[8]HKUST (Guangzhou), Guangzhou, China
[9]Department of Computer Science and Technology, Peking University, Beijing, China
[10]These authors contributed equally
*Correspondence: wangyuanzhuo@ict.ac.cn (Y.W.); guojian@idea.edu.cn (J.G.)

Received: September 25, 2025; Accepted: December 31, 2025; https://doi.org/10.1016/j.xinn.2025.101253

Citation: Gu J., Jiang X., Shi Z., et al., (2026). A survey on LLM-as-a-judge. The Innovation **7(6)**, 101253.

Accurate and consistent evaluation is crucial for decision-making across numerous fields, yet it remains challenging due to inherent subjectivity, variability, and scale. Large language models (LLMs) have achieved remarkable success, leading to "LLM-as-a-judge," where LLMs serve as evaluators for complex tasks. With their ability to process diverse data types and provide scalable assessments, LLMs present a compelling alternative to traditional expert-driven evaluations. However, ensuring the reliability of LLM-as-a-judge systems remains a significant challenge requiring careful design and standardization. This paper provides a comprehensive survey of LLM-as-a-judge, offering a formal definition and detailed classification while addressing the core question of how to build reliable LLM-as-a-judge systems. We explore strategies to enhance reliability, including improving consistency, mitigating biases, and adapting to diverse scenarios. We propose methodologies for evaluating reliability, supported by a novel benchmark. To advance development and deployment, we discuss practical applications, challenges, and future directions. Our contributions span multiple levels: we establish conceptual boundaries, reorganize fragmented literature into a unified framework, and propose a reliability-oriented benchmark. We articulate a forward-looking research agenda, offering theoretical foundations and practical guidance for constructing reliable and trustworthy LLM-as-a-judge systems.

INTRODUCTION

to scale, and susceptible to inconsistency. On the other hand, objective assessment methods, such as automatic metrics, offer strong scalability and consistency. For example, tools such as BLEU[19] or ROUGE[20] can rapidly evaluate machine-generated translations or summaries against reference texts without human intervention. However, these metrics, which heavily rely on surface-level lexical overlap, often fail to capture deeper nuances, resulting in poor performance in tasks such as story generation or instructional texts.[21] As a solution to this persistent dilemma, LLM-as-a-judge has emerged as a promising idea to combine the strengths of the above two evaluation methods. Recent studies have shown that this idea can merge the scalability of automatic methods with the detailed, context-sensitive reasoning found in expert judgments.[15,22–25] Moreover, LLMs may become sufficiently flexible to handle multimodal inputs[26] under appropriate prompt learning or fine-tuning.[27] These advantages suggest that the LLM-as-a-judge approach could serve as a novel and broadly applicable paradigm for addressing complex and open-ended evaluation problems.

LLM-as-a-judge holds significant potential as a scalable and adaptable evaluation framework compared to the aforementioned two traditional methods.[28] However, its widespread adoption is hindered by two key challenges. The first challenge lies in the absence of a systematic review, which highlights the lack of formal definitions, fragmented understanding, and inconsistent usage practices in the relevant studies. As a result, researchers and practitioners struggle to fully understand and apply effectively. The second challenge concerns reliability,[29] as merely employing LLM-as-a-judge does not ensure accurate evaluations aligned with established standards. These challenges emphasize the need for a deeper assessment of the outputs generated by

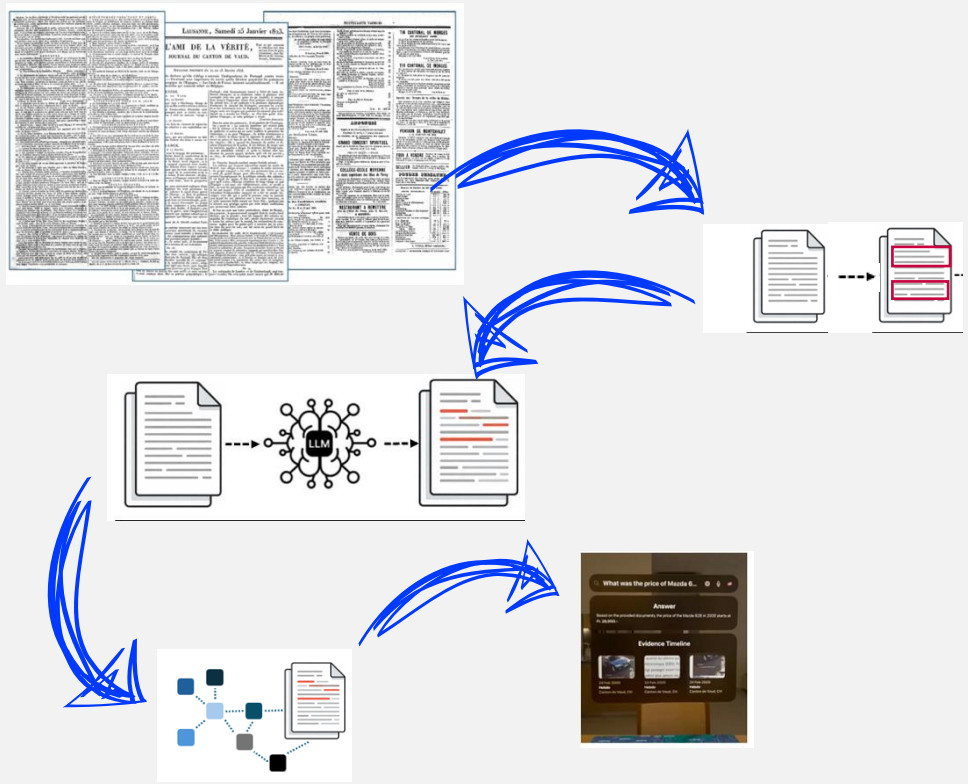
Definitions| What is LLM as a judge

Core idea

Simulate human Judgment using LLMs

- **Lower costs** in both time and money
- **Facilitate experimentation** in different situations

Concrete examples



Motivation| LLM as a judge can benefit everyone in our department (and beyond)

In many instances...

Analyse user studies (e.g., interviews, user forms)

Algorithms/pipelines (e.g., Information retrieval, pipelines)

Bias understanding (e.g., how does one person's perspective differ from someone else's)

Information retrieval (e.g., company data, user posts)

...we can use LLM as a judge

→ **Select**
(e.g., Follow up questions, insights)

→ **Score**
(e.g., Performance metrics)

→ **Classify**
(e.g., Biased/Unbiased, Propaganda...)

→ **Compare**
(e.g., Which document is most relevant, is it found or not...)

Summary| The article Explores what LLM as a judge is and different steps in an implementation

Survey design

In this Survey, the authors explore…

- **How to use**
- **How to improve**
- **How to evaluate**
- **Applications**
- **Challenges**
- **Future work**

Summary| How does the paper approach the lifecycle of LLM as a judge

Findings along three key steps

1. **How to use**
   Getting started
2. **How to improve**
   Common tricks
3. **How to evaluate**
   Validation

In order to successfully implement an LLM as a judge one needs to…

Identify the target to measure *(model, data, agents, reasoning?) Define clear* ***reliability metrics****, select a* ***model*** *carefully, as well as a* ***prompt/context.***

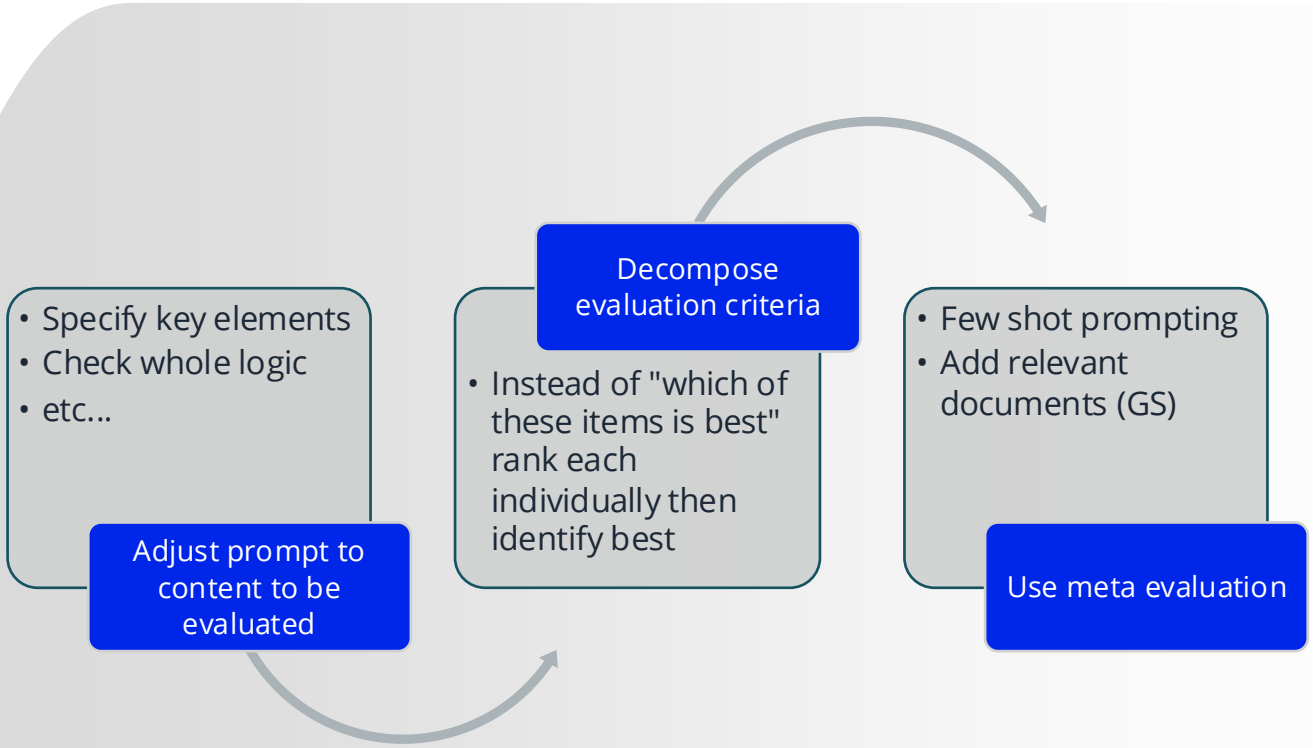
Structure your experiment around LLM as a judge, expect an iterative effort

Gu et al. (2026, p. 2-8)

Summary| How does the paper approach the lifecycle of LLM as a judge

Findings along three key steps

1. **How to use**
   Getting started
2. **How to improve**
   Common tricks
3. **How to evaluate**
   Validation



Gu et al. (2026, p. 8-12)

Summary| How does the paper approach the lifecycle of LLM as a judge

Findings along three key steps

1. **How to use**
   Getting started
2. **How to improve**
   Common tricks
3. **How to evaluate**
   Validation

Agreement with human judge	$\frac{\Sigma_{I \in D} I(S_{LLM} = S_{Human})}{D}$	If the LLM evaluation matches a human's evaluation
Bias Judgment specific biases	$\left\lvert \frac{\Sigma_{I \in D} I(S_{LLM} - S_{Human})}{D} \right\rvert > \Delta$	If the LLM evaluation consistently deviates from a neutral evaluation
Adversarial robustness	$S_{LLM} = S_{Corrupted\ LLM} \pm \Delta$	If an adversary can meaningfully corrupt one's evaluations

Gu et al. (2026, p. 13-14)

Summary| How does the paper approach the Applications and Challenges

Creative applications

evaluate viability of creative ideas

Bias amplification

Impact on creative and diverse outputs

Critical decision making

reviewing patient data or approving financial decisions

Lack of accountability

Risk of adversarial input (high damage potential)

Review specific applications

LLM reviewer #2

Input sensitivity

Lack of domain specific knowledge

Gu et al. (2026, p. 16-22)

Strenghts| What this paper did well

Approaching the subject

The authors create a highly accessible roadmap for both academics and practitioners.

AI bias analysis

It distinctly separates "task-agnostic biases" from "judgment-specific biases"

Formalization

Formal definition of the "LLM-as-a-judge" process, encapsulated as

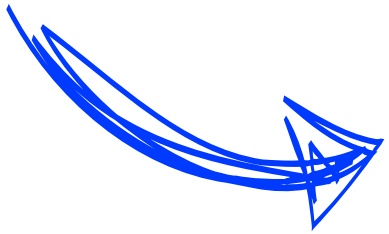
$$\mathcal{R} \leftarrow f_{\mathbb{R}}\left(P_{\mathcal{CMM}}, x, C\right)$$

Limitations| The paper needs to be interpreted considering the key limitations in the definition, methodology, and presentation

	Definition	Methodology	Meta-evaluation
Context	The paper aims to standardise the terminology around LLM-as-a-Judge	The authors compare performance across models/setups	The paper aims to “evaluate the evaluator”
Limitation	**Creating confusion between ideas and effective implementations**	**The comparison utilises closed source obsolete model's vs low tier open-source models**	**The paper struggles to provide a scalable solution that does not fall back on human annotation.**
Examples	Multi-LLM vs LLM voting Prompt train data/in context learning vs Few shot prompting/prompt engineering	Use of GPT-3.5 (2022) and Qwen 2.5 7b (2024)	

Relevancy| In my dissertation, the article is relevant in two key ways

Using the findings…



1

…to **better evaluate our pipeline** *across different steps, and allow for a faster and more reliable iteration process*

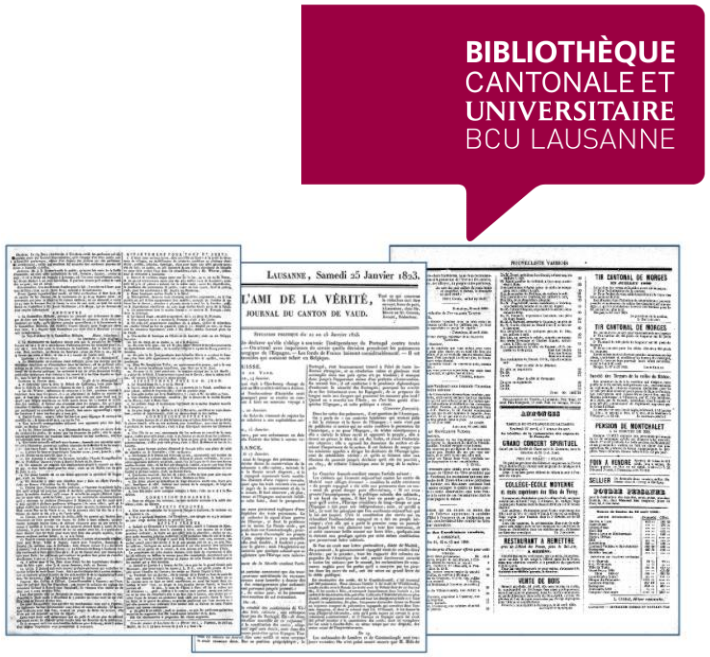
2

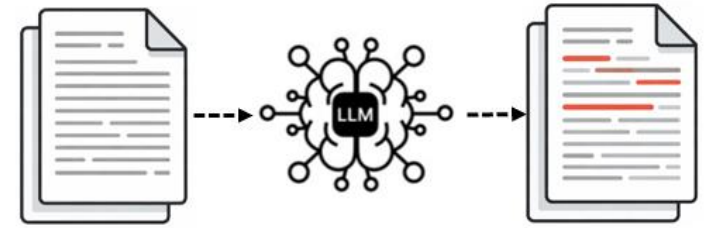
…to **align our evaluations** *with academic and practitioner metrics more thoroughly*

THANK YOU!

Unil.

Teaser| A preview of what my research is on





1) LLM-driven text data sanitization

2) RAG-based document retrieval (chatbot)