Artificial Intelligence and Language Models

ACS Publishing

Editorial

LLM-as-a-Judge in Evaluation-Centric AI: Trends and Challenges

Aytug Onan[1]*, Arbi Haza Nasution[2]

[1] Department of Computer Engineering, Faculty of Engineering, Izmir Institute of Technology, Türkiye
[2] Department of Informatics Engineering, Universitas Islam Riau, Indonesia
* Correspondence: aytugonan@iyte.edu.tr;

Abstract: Large Language Models (LLMs) have undergone a rapid transformation from architectural innovations to widely deployed intelligent systems. This editorial traces the evolution of LLMs through a structured timeline, beginning with the introduction of the Transformer architecture, which enabled scalable attention-based learning across diverse language tasks. This foundation was extended through generative pre-training paradigms, leading to increasingly capable models such as GPT-3, which demonstrated few-shot learning at scale. Subsequent developments addressed key limitations of parametric models, particularly their static knowledge boundaries, through the introduction of Retrieval-Augmented Generation (RAG), enabling dynamic integration of external information during inference. The emergence of instruction tuning and reinforcement learning from human feedback further enabled alignment with human intent, culminating in the deployment of conversational systems such as ChatGPT. More recently, the field has entered a new phase characterized by the rise of open-weight LLM ecosystems and evaluation-centric methodologies. In particular, the introduction of LLM-as-a-judge has redefined evaluation by enabling scalable, model-based assessment of generated outputs, while LLM-based annotation and multi-agent evaluation frameworks have further expanded the role of LLMs from generators to evaluators. This progression reflects a fundamental shift in LLM research, from model-centric development toward evaluation, benchmarking, robustness, and real-world deployment. Despite this shift, existing publication venues remain largely focused on model design, leaving a gap for evaluation-driven research. To address this need, we introduce *Artificial Intelligence and Language Models (AILM)*, an open-access journal dedicated to advancing research on LLM evaluation, benchmarking, LLM-as-a-judge frameworks, retrieval-augmented systems, and low-resource language applications. AILM aims to provide a platform for the next phase of AI research, where reliability, reproducibility, and real-world impact are central concerns.

Keywords: Large Language Models (LLMs), LLM-as-a-judge, Benchmarking, Evaluation of Language Models, Retrieval-Augmented Generation (RAG), AI Robustness, Low-Resource Languages, Automated Annotation, AI Deployment, Natural Language Processing (NLP)

Citation: Aytug Onan, Arbi Haza Nasution. LLM-as-a-Judge in Evaluation-Centric AI: Trends and Challenges. *Artificial Intelligence and Language Models* 1–13. https://doi.org/

Received: 25 March 2026
Accepted: 25 March 2026
Published: 25 March 2026

1. Introduction

Large Language Models (LLMs) have evolved rapidly over the past decade, transforming from architectural innovations into widely deployed socio-technical systems. This evolution is not merely a progression in model size or performance, but a fundamental shift in how artificial intelligence systems are designed, evaluated, and deployed. While early work emphasized architectural innovation and performance improvements, recent developments highlight the importance of evaluation, robustness, and real-world applicability.

Figure 1 presents a high-level view of the evolution of LLM research, illustrating the transition from architectural innovation to evaluation-centric AI. Figure 2 presents a structured view of this evolution, highlighting a transition from model-centric research toward evaluation-centric and deployment-aware paradigms. Early work focused on

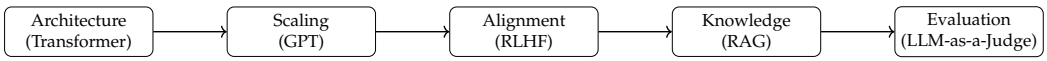

Figure 1. Graphical abstract illustrating the evolution of LLM research from architectural innovation to evaluation-centric AI.

architectural design and pre-training strategies, followed by scaling and alignment techniques that enabled general-purpose conversational systems. More recent developments emphasize retrieval-based knowledge integration, automated evaluation frameworks, and domain-specific benchmarking.

We argue that this evolution can be understood as a transition across five distinct phases: (1) architectural innovation, (2) scaling and pre-training, (3) alignment and deployment, (4) knowledge integration, and (5) evaluation and benchmarking. This five-phase framework provides a unifying perspective that situates recent advances, such as LLM-as-a-judge, within a broader research trajectory in which evaluation, benchmarking, robustness, and reproducibility are no longer secondary concerns but central scientific challenges.

This five-phase framework serves as the central narrative of this paper, positioning LLM-as-a-judge and evaluation-centric methodologies as the defining characteristics of the current phase of AI research.

2. From Architecture to Evaluation: A Timeline Perspective

To contextualize this transition, we first examine the historical evolution of LLMs through the key milestones illustrated in Figure 2. Rather than treating these developments as isolated breakthroughs, we interpret them as stages in a broader transition from model construction to evaluation and deployment.

2.1. Stage 1: Architectural Foundations (2017–2019)

The first stage established the fundamental computational paradigm for modern LLMs, emphasizing scalability and parallelization. The introduction of the Transformer architecture [1] enabled scalable attention-based modeling, allowing models to learn multiple language tasks by focusing on relevant input relationships. This innovation was followed by GPT-style generative pre-training [2,3], which reframed NLP tasks as conditional generation problems and introduced the concept of multitask learning through language modeling.

2.2. Stage 2: Scaling and Knowledge Integration (2020)

This stage demonstrated that model scale, data, and compute are primary drivers of emergent capabilities. The release of GPT-3 [4] demonstrated the power of large-scale models and few-shot learning. In parallel, Retrieval-Augmented Generation (RAG) [5] addressed a critical limitation of parametric models by enabling dynamic retrieval of external knowledge during inference. This marked a transition toward hybrid systems combining parametric and non-parametric knowledge.

A recurrent failure mode of parametric-only LLMs is that knowledge is bounded by training data and post-training; updates are non-trivial and provenance is unclear. RAG reframes generation as a probabilistic composition of a *retriever* and a *generator* [5]. In RAG, retrieval candidates are treated as latent variables and marginalised (approximately) in the generation process. In practice, modern RAG pipelines often include:

- **Indexing:** chunking, embedding, and storing passages (or structured nodes) in an index.
- **Retrieval:** sparse (e.g., BM25) or dense retrieval; DPR is a canonical dense bi-encoder retriever [32].
- **Grounded generation:** concatenating retrieved context into a prompt or conditioning window.
- **Verification:** optional re-ranking, citation extraction, and faithfulness checks.

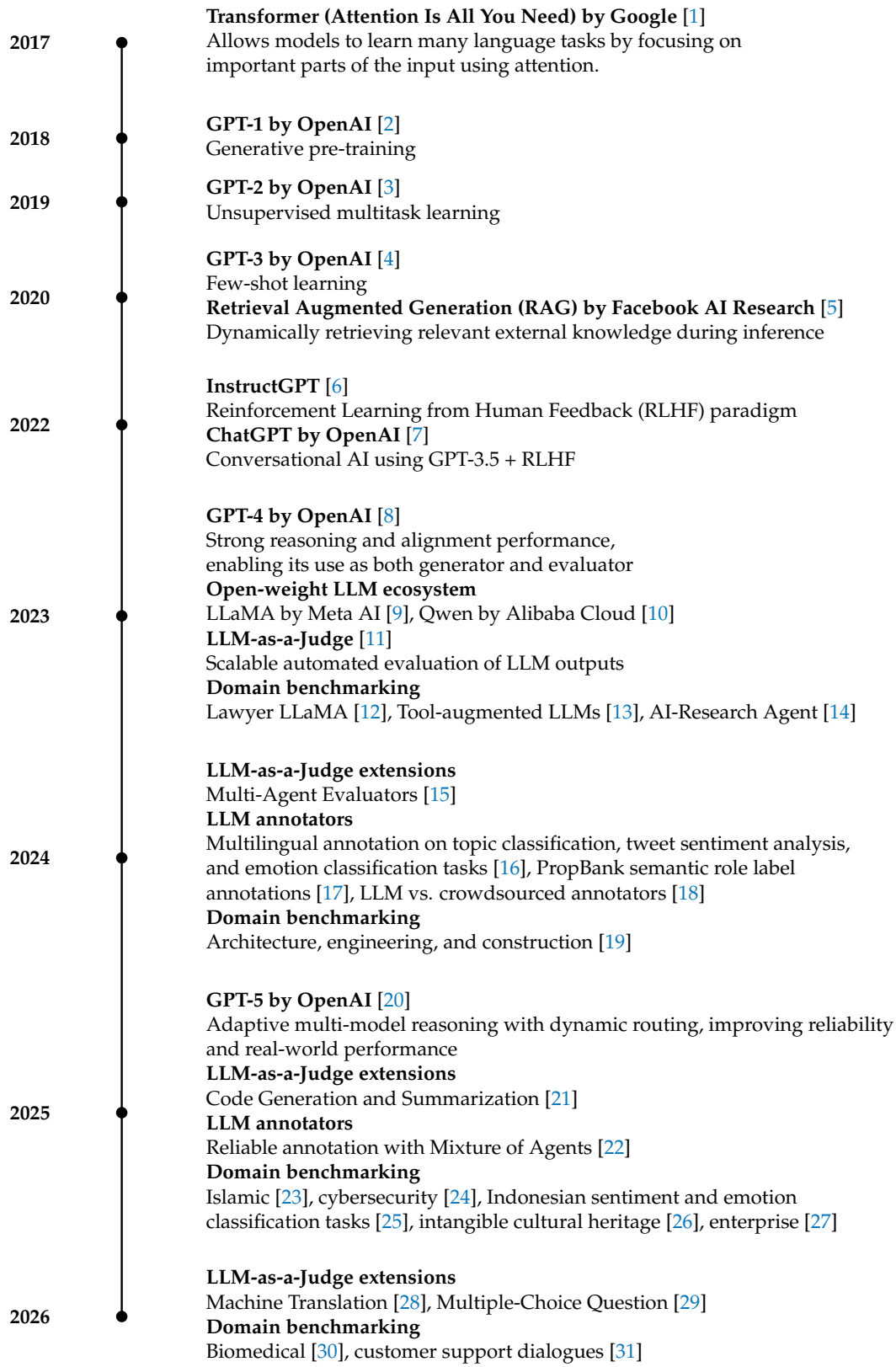

Figure 2. Integrated vertical timeline of large language model evolution, combining architectural advances (GPT series), training paradigms (RLHF), and evaluation frameworks (RAG, LLM-as-a-judge, benchmarking). The figure highlights the transition from model-centric development toward evaluation-centric AI research.

RAG does not "solve hallucination" categorically; rather, it changes its geometry: failures can come from retrieval misses, chunking artefacts, prompt-context conflicts, or model unfaithfulness to evidence [33]. For AILM, RAG is therefore not merely an application area; it is an evaluation challenge: we must measure retrieval quality, grounding faithfulness, and end-to-end accuracy under realistic constraints.

2.3. Stage 3: Alignment and Deployment (2022)

This stage marked the transition from experimental models to real-world AI systems deployed at scale. Instruction tuning and reinforcement learning from human feedback (RLHF) [6] enabled models to better align with user intent. The release of ChatGPT [7] demonstrated the practical deployment of LLMs as conversational systems, bringing LLM capabilities into widespread real-world use.

Scaling alone does not guarantee helpful behaviour. Instruction-following and RLHF-based post-training demonstrated that models can be tuned to follow user intent more reliably, even without increasing parameter count [6]. ChatGPT's public research preview then made this alignment paradigm accessible at global scale [7]. In the editorial timeline (Figure 2), we treat 2018 as the start of the GPT-style trajectory, and 2022 as the inflection point for mainstream conversational deployment.

2.4. Stage 4: Open Ecosystem and Evaluation Paradigms (2023)

The fourth stage represents a shift toward hybrid systems that combine parametric learning with dynamic retrieval. The emergence of open-weight LLMs such as LLaMA [9] and Qwen [10] enabled reproducible research and accelerated benchmarking. Open-weight LLMs reshaped research incentives. With LLaMA, Meta argued that strong foundation models can be trained on public data and released for research, enabling replication and comparative study at a speed rare in closed-model ecosystems [9]. From an editorial perspective, this matters because *deployment constraints become part of the scientific problem*. If results depend on a transient API endpoint, hidden system prompts, or silent model updates, then evaluation becomes non-reproducible by design. Open stacks do not solve all reproducibility issues, but they make them visible and, therefore, researchable.

High-quality labels are expensive. Expert annotation offers depth and domain sensitivity but scales poorly. Crowdsourcing (e.g., Mechanical Turk) improves speed and cost but raises quality-control and variance issues; classic NLP work demonstrated that many tasks can be annotated reasonably well by non-experts with appropriate design and aggregation [34]. This cost–quality boundary is one reason model-mediated evaluation became attractive. The introduction of LLM-as-a-judge [11] marked a fundamental shift in evaluation methodology, allowing models to assess and compare the outputs of other models. LLM-as-a-judge formalises the use of strong models as scalable proxies for human preference, validated against controlled and crowdsourced human judgements [11]. Yet this paradigm introduces new failure modes: position/verbosity bias, self-enhancement bias, and sensitivity to prompt framing. G-Eval further illustrates that even when LLM evaluators correlate better with humans than classical metrics, evaluators can be biased toward LLM-like output styles [35].

2.5. Stage 5: Automated Evaluation and Benchmarking (2024–2026)

In this last stage, evaluation becomes a first-class component of AI systems rather than a post-hoc analysis step. Recent developments extend evaluation capabilities through LLM-based annotation [16], multi-agent evaluation frameworks [15,22], and domain-specific benchmarking across diverse fields, including Islamic [23], cybersecurity [24], Indonesian sentiment and emotion classification tasks [25], intangible cultural heritage [26], and enterprise [27].

This progression highlights a clear shift in research focus, from building increasingly powerful models toward understanding how to evaluate, compare, and deploy them

effectively. This shift motivates the need for new evaluation paradigms, discussed in the following sections.

3. Evaluation as the New Core of LLM Research

As LLM capabilities have expanded, traditional evaluation methods have become increasingly insufficient. This has led to a transition from metric-based evaluation toward more flexible, scalable, and context-aware approaches.

The timeline highlights a clear shift from model development toward evaluation-centric research. This shift is driven by three key factors. First, LLM outputs are increasingly open-ended, making traditional reference-based metrics insufficient. Recent work on multi-agent evaluation [22] suggests that reliability can be improved by aggregating multiple model outputs, highlighting a shift toward ensemble-based evaluation strategies. Second, the cost of human annotation necessitates scalable alternatives such as LLM-based evaluation. Recent work shows LLM-based annotation reduces labeling costs but requires careful validation [16]. Third, real-world deployment requires robust and reproducible evaluation frameworks. LLM-as-a-judge [11] represents a scalable solution to evaluation, but introduces new challenges that define the next phase of AI:

- Evaluator bias and consistency
- Sensitivity to prompt design
- Alignment between human and model judgments
- Robustness under adversarial or ambiguous inputs
- Cross-domain and multilingual generalization
- Reproducibility under model updates and system drift

These challenges indicate that evaluation is no longer a supporting task, but a complex and evolving research domain requiring dedicated methodologies.

4. LLM-as-a-Judge: A New Evaluation Paradigm

A foundational study of this paradigm is presented by Zheng et al. [11], who systematically investigate the use of LLMs as evaluators for open-ended tasks. Their work highlights the limitations of traditional benchmarks in capturing human preferences and proposes LLM-based evaluation as a scalable alternative. To validate this approach, they introduce two benchmarks: MT-Bench, a multi-turn question set designed to assess conversational capabilities, and Chatbot Arena, a crowdsourced platform for pairwise comparison of model outputs. Their findings show that strong LLM evaluators, such as GPT-4, achieve over 80% agreement with human judgments, a level comparable to inter-human agreement. This result demonstrates that LLM-as-a-judge can serve as an effective proxy for human evaluation, while significantly reducing cost and improving scalability.

The concept of LLM-as-a-judge refers to the use of large language models (LLMs) as evaluators for assessing the quality of outputs in natural language processing (NLP) tasks. This paradigm has gained significant attention as it enables nuanced, scalable, and context-aware evaluation that goes beyond traditional metrics such as BLEU, ROUGE, or accuracy, which often fail to capture the complexity of language generation tasks [36].

4.1. Applications of LLM-as-a-Judge

LLM-as-a-judge has been applied across a wide range of NLP tasks. In natural language generation (NLG), it is used to evaluate summarization, question answering, story generation, and machine translation, where multiple valid outputs exist and reference-based metrics are insufficient [37]. In code-related tasks, LLMs are used to assess code generation and summarization, although even advanced models may misjudge correctness in complex scenarios [38]. Furthermore, LLM-based evaluators have been applied in multilingual settings, enabling large-scale evaluation across languages, although biases and inconsistencies remain challenges [39].

4.2. Strengths of LLM-as-a-Judge

One of the key strengths of LLM-as-a-judge is its ability to provide nuanced evaluation. Unlike traditional metrics, LLMs can assess qualitative aspects such as coherence, tone, factuality, and hallucination, enabling more meaningful evaluation of open-ended tasks [36].

Another advantage is scalability. LLM-based evaluation significantly reduces reliance on human annotators, enabling large-scale benchmarking and continuous evaluation in deployment environments [38].

In addition, LLMs offer flexibility in evaluation protocols, supporting approaches such as pairwise comparison, ranking, and critique-based evaluation, as demonstrated in generative judge frameworks [40].

4.3. Challenges and Limitations

In addition to demonstrating effectiveness, Zheng et al. [11] identify several important limitations of LLM-based evaluation, including position bias, verbosity bias, and self-enhancement bias. These biases highlight the need for careful prompt design, calibration strategies, and standardized evaluation protocols to ensure reliable and fair assessments.

A major concern is bias, including self-referential bias, where models tend to favor outputs similar to their own generation patterns, leading to skewed evaluations [36,39]. Another limitation is inconsistency. LLM-based evaluations may vary across prompts, temperature settings, or model versions, raising concerns about reproducibility and stability of results [38]. Task-specific limitations also exist. Smaller models often struggle with complex evaluation tasks, and even advanced models may fail in specialized domains such as code generation [38].

Finally, multilingual evaluation introduces additional challenges, including bias toward high-resource languages and the need for calibration with human judgments to ensure fairness and accuracy [39].

4.4. Comparison with Traditional Evaluation Metrics

Compared to traditional metrics, LLM-as-a-judge provides a more flexible and semantically rich evaluation framework. While traditional metrics are efficient and reproducible, they are limited to surface-level comparisons and often fail in open-ended tasks. In contrast, LLM-based evaluation captures deeper semantic and contextual aspects but introduces challenges in reliability and bias. This trade-off highlights the need for hybrid evaluation approaches that combine both methods [36].

4.5. Toward Reliable and Scalable Evaluation Frameworks

Future research should focus on developing standardized evaluation protocols, improving calibration techniques to align LLM judgments with human evaluation, and designing hybrid frameworks that integrate LLM-based evaluation with traditional metrics and human oversight [41].

4.6. From Evaluation Tool to System Component

Importantly, LLM-as-a-judge is evolving from a standalone evaluation tool into a core component of modern AI systems. As illustrated in Figure 3, evaluation is increasingly integrated into generation pipelines, forming closed-loop systems that continuously assess and improve outputs. This transformation reinforces the emergence of evaluation-centric AI, where evaluation is embedded directly into system design rather than applied as a post-hoc process.

5. From Model-Centric to Evaluation-Centric AI

We define evaluation-centric AI as a paradigm in which the primary research focus shifts from improving model performance to ensuring reliability, robustness, and trustworthiness.

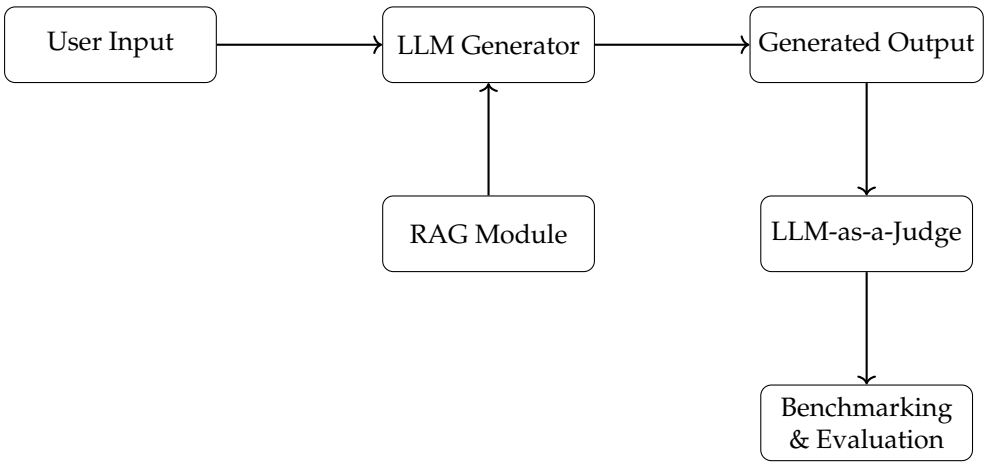

Figure 3. Evaluation-centric LLM pipeline integrating generation, retrieval, and automated evaluation. The figure illustrates how modern LLM systems extend beyond generation to include retrieval-augmented knowledge and LLM-based evaluation, forming a closed-loop system for benchmarking and deployment.

Figure 3 and Table 1 together illustrate the transition toward evaluation-centric AI. While traditional systems focused on improving model performance in isolation, modern LLM systems integrate generation, retrieval, and evaluation into a unified pipeline. This shift highlights the increasing importance of benchmarking, automated evaluation, and system-level design in contemporary AI research.

This paradigm is characterized by:

- Benchmarking across domains and tasks
- Automated evaluation frameworks (e.g., LLM-as-a-judge)
- Integration of evaluation into system design
- Focus on real-world deployment

This transition fundamentally changes how AI systems are developed, evaluated, and deployed.

6. A Taxonomy of LLM Evaluation Methods

To better understand the emerging landscape of evaluation-centric AI, we propose a taxonomy of LLM evaluation methods based on their underlying evaluation mechanisms and level of automation. This taxonomy provides a structured view of how evaluation has evolved from traditional metrics to model-based and system-level approaches.

6.1. Metric-Based Evaluation

Metric-based evaluation relies on predefined quantitative measures, such as accuracy, BLEU, ROUGE, and F1 score. These metrics are widely used for tasks with well-defined ground truth, including classification and machine translation. However, they are limited in capturing semantic quality, reasoning ability, and contextual appropriateness in open-ended tasks.

6.2. Human-Based Evaluation

Human evaluation remains the gold standard for assessing language model outputs, particularly for tasks involving creativity, reasoning, and contextual understanding. Human annotators can evaluate fluency, coherence, and factual correctness. However, this approach is expensive, time-consuming, and difficult to scale, especially for large datasets and continuous deployment settings.

Table 1. From Model-Centric to Evaluation-Centric AI: A Paradigm Shift in LLM Research

Aspect	Model-Centric AI	Evaluation-Centric AI
Primary Goal	Improve model performance (accuracy, perplexity)	Ensure reliability, robustness, and real-world usability
Core Focus	Model architecture and training	Evaluation, benchmarking, and deployment
Key Methods	Transformer architectures, scaling laws, pre-training	LLM-as-a-judge, benchmarking, multi-agent evaluation
Evaluation Approach	Static metrics (BLEU, accuracy, F1)	Dynamic evaluation (LLM-based judging, human-AI alignment)
Data Dependency	Large labeled datasets	Combination of labeled data, retrieval, and synthetic evaluation
System Design	Single-model systems	Multi-component systems (LLM + RAG + evaluator)
Challenges	Model optimization and generalization	Bias, consistency, reproducibility, evaluation reliability
Application Scope	Benchmark datasets	Real-world deployment across domains
Representative Phase	GPT-1 to GPT-3 era	GPT-4, LLM-as-a-judge, benchmarking era

6.3. LLM-as-a-Judge Evaluation

LLM-as-a-judge represents a paradigm shift in evaluation, where language models are used to assess the outputs of other models. This approach enables scalable, automated evaluation across diverse tasks, including reasoning, dialogue, and summarization. Despite its advantages, challenges remain in ensuring consistency, reducing bias, and aligning with human judgment.

6.4. Reference-Free Evaluation

Reference-free evaluation methods assess outputs without relying on ground truth labels. These approaches are particularly useful for open-ended generation tasks, where multiple valid outputs may exist. LLM-based evaluators often operate in this setting by comparing candidate responses or scoring outputs based on qualitative criteria.

6.5. Retrieval-Augmented Evaluation

Retrieval-augmented evaluation integrates external knowledge sources to verify factual correctness. By leveraging retrieval mechanisms, evaluators can cross-check generated outputs against trusted documents, improving the reliability of evaluation in knowledge-intensive tasks.

6.6. Multi-Agent and Ensemble Evaluation

Multi-agent evaluation frameworks combine multiple evaluators, either human or LLM-based, to improve reliability and reduce bias. Ensemble approaches aggregate judgments from multiple models, enabling more robust and consistent evaluation outcomes.

6.7. System-Level and Deployment Evaluation

System-level evaluation focuses on the performance of LLMs within real-world applications, considering factors such as latency, user satisfaction, robustness, and safety. This type of evaluation extends beyond isolated model outputs to assess end-to-end system behavior.

6.8. Summary of the Taxonomy

Table 2 summarizes the proposed taxonomy of LLM evaluation methods. This taxonomy highlights the transition from static, metric-based evaluation toward dynamic,

Table 2. Taxonomy of LLM Evaluation Methods

Category	Description	Strengths and Limitations
Metric-Based	Quantitative metrics such as accuracy, BLEU, ROUGE, F1	Efficient and reproducible, but limited for open-ended tasks
Human-Based	Evaluation by human annotators	High quality and interpretability, but costly and non-scalable
LLM-as-a-Judge	LLMs evaluating model outputs	Scalable and flexible, but subject to bias and inconsistency
Reference-Free	Evaluation without ground truth labels	Suitable for generative tasks, but harder to standardize
Retrieval-Augmented	Uses external knowledge to verify outputs	Improves factual accuracy, but depends on retrieval quality
Multi-Agent / Ensemble	Combines multiple evaluators	More robust and reliable, but computationally expensive
System-Level	Evaluates full deployment systems	Realistic and application-driven, but complex to design

model-based, and system-level approaches, reflecting the broader shift toward evaluation-centric AI.

This taxonomy provides a unifying framework for understanding the diverse evaluation methodologies in LLM research, and serves as a foundation for future work in benchmarking, automated evaluation, and system-level assessment.

7. A Practical Checklist for LLM Evaluation

To support reproducibility and standardization in evaluation-centric AI, we propose a practical checklist for LLM evaluation. This checklist summarizes key considerations that researchers and practitioners should address when designing and reporting evaluation experiments.

Table 3 outlines a structured checklist for LLM evaluation that spans the full lifecycle of evaluation design, from task formulation to deployment considerations. Unlike traditional evaluation protocols that focus primarily on performance metrics, this checklist emphasizes a holistic approach that integrates methodological rigor, robustness, and real-world applicability.

The first group of criteria, including *Task Definition*, *Evaluation Method*, and *Ground Truth*, ensures that the evaluation problem is well-specified and aligned with the intended application. Clearly defining the task and selecting appropriate evaluation methods are critical for avoiding misleading conclusions, particularly in open-ended generation tasks where multiple valid outputs may exist.

The second group, consisting of *Model Setup* and *Reproducibility*, addresses experimental transparency. In the context of LLMs, where performance can be highly sensitive to prompt design, model versioning, and parameter settings, reproducibility is essential for ensuring that results can be independently verified and compared across studies.

The third group focuses on reliability and trustworthiness, including *Bias and Fairness*, *Robustness*, and *Consistency*. These aspects capture the behavior of models under diverse and potentially adversarial conditions, as well as their stability across repeated evaluations. Addressing these factors is particularly important for deployment in high-stakes domains.

The final group, *Human Alignment* and *Deployment Readiness*, extends evaluation beyond laboratory settings to real-world applications. Alignment with human judgment ensures that model outputs are meaningful and acceptable to users, while deployment considerations such as latency, cost, and user experience determine the practical feasibility of LLM systems.

Taken together, this checklist reflects a shift from performance-centric evaluation toward system-level evaluation, where reliability, reproducibility, and usability are equally

Table 3. Checklist for LLM Evaluation and Benchmarking

Evaluation Aspect	Key Questions
Task Definition	Is the task clearly defined? Does it reflect real-world usage?
Evaluation Method	Are appropriate evaluation methods selected (metric-based, LLM-as-a-judge, human)?
Ground Truth	Is ground truth available, or is evaluation reference-free?
Model Setup	Are model versions, prompts, and parameters clearly specified?
Reproducibility	Can the experiment be reproduced with available data and configuration?
Bias and Fairness	Are potential biases evaluated across domains and languages?
Robustness	Is the model tested under adversarial or noisy inputs?
Consistency	Are evaluation results stable across multiple runs or prompts?
Human Alignment	Do model judgments align with human evaluations?
Deployment Readiness	Are latency, cost, and user experience considered?

important as accuracy. As such, it can serve as a standardized framework for future benchmarking studies and evaluation protocols in LLM research.

This checklist provides a standardized reference for designing rigorous and reproducible evaluation pipelines, and can serve as a baseline for future benchmarking studies. We recommend that future LLM evaluation studies explicitly report these dimensions to improve comparability, transparency, and scientific rigor across the field.

8. Future Directions and Open Challenges

The transition toward evaluation-centric AI introduces a new set of research challenges that will define the next phase of large language model development. While current advances demonstrate the feasibility of automated evaluation and deployment, several critical open problems remain unresolved.

8.1. Reliability and Consistency of LLM-as-a-Judge

Although LLM-as-a-judge enables scalable evaluation, its reliability remains an open question. Models may exhibit variability across prompts, sensitivity to phrasing, and inconsistencies across evaluation runs. Future research should focus on improving calibration, stability, and agreement with human judgments, as well as developing standardized evaluation protocols.

8.2. Bias and Alignment in Automated Evaluation

LLM-based evaluators inherit biases from their training data and alignment procedures. This raises concerns about fairness, cultural bias, and domain-specific misjudgment. Addressing these issues requires systematic bias auditing, cross-cultural evaluation benchmarks, and hybrid human-AI evaluation frameworks.

8.3. Benchmarking in Dynamic and Open-Ended Tasks

Traditional benchmarks are static and often fail to capture the complexity of real-world applications. Future benchmarking frameworks must support open-ended tasks, dynamic data, and evolving evaluation criteria, enabling more realistic assessment of LLM performance in deployment settings.

8.4. Evaluation in Low-Resource and Multilingual Contexts

Most evaluation frameworks are designed for high-resource languages, limiting their applicability in global contexts. Developing reliable evaluation methodologies for low-resource languages remains a critical challenge, particularly for tasks involving cultural nuance and domain-specific knowledge.

8.5. Integration of Evaluation into System Design

Evaluation is increasingly becoming an integral component of LLM systems rather than a post-hoc process. Future research should explore architectures that embed evaluation directly into generation pipelines, enabling adaptive systems that continuously improve based on feedback.

8.6. Reproducibility and Model Drift

Frequent updates to LLMs introduce challenges in reproducibility and longitudinal evaluation. Model drift can affect benchmark results and invalidate prior comparisons. Establishing reproducible evaluation protocols and version-controlled benchmarks is essential for maintaining scientific rigor.

8.7. Toward Unified Evaluation Frameworks

A key open problem is the lack of unified frameworks that integrate multiple evaluation dimensions, including accuracy, reasoning quality, robustness, and user satisfaction. Future work should aim to develop comprehensive evaluation systems that combine automatic metrics, LLM-based judgments, and human feedback.

These challenges highlight that evaluation is not a solved problem but an emerging research frontier. Addressing these issues will be critical for ensuring that LLMs are reliable, trustworthy, and effective in real-world applications.

9. Why a Dedicated Journal is Needed

Figure 2 illustrates that LLM research has evolved beyond model construction into a broader ecosystem involving evaluation, benchmarking, and deployment. However, existing journals primarily focus on algorithmic and architectural contributions, with limited emphasis on systematic evaluation methodologies.

This creates a gap in the research landscape. Key topics such as LLM-as-a-judge, domain-specific benchmarking, low-resource evaluation, and reproducibility often lack a dedicated venue that treats evaluation as a primary contribution.

The journal *Artificial Intelligence and Language Models (AILM)* is established to address this gap. It provides a platform for research that prioritizes:

- Benchmarking and evaluation of LLMs
- Robustness and adversarial analysis
- Retrieval-augmented systems
- LLM-based evaluation and annotation
- Low-resource and multilingual NLP
- Real-world AI deployment and reproducibility

By focusing on these areas, AILM aligns with the current trajectory of LLM research and supports the development of trustworthy AI systems.

10. Conclusion

The evolution of large language models reflects a transition from architectural innovation to evaluation-centric research. While early work focused on building increasingly powerful models, recent developments emphasize the need for reliable, robust, context-aware evaluation and applicability in real-world settings. Among these developments, LLM-as-a-judge emerges as a defining paradigm that fundamentally reshapes how language models are evaluated. By enabling scalable, flexible, and semantically rich assessment of open-ended tasks, LLM-based evaluation provides a practical alternative to traditional metrics and costly human annotation, while introducing new challenges related to bias, consistency, and reproducibility.

The next phase of LLM research will therefore be defined not only by larger or more capable models, but by advances in evaluation frameworks, particularly those centered on LLM-as-a-judge, as well as stronger reproducibility practices and more trustworthy

deployment strategies. LLM-as-a-judge is likely to become a standard component of future AI systems, where evaluation is integrated directly into generation pipelines rather than treated as a post-hoc process.

Artificial Intelligence and Language Models (AILM) is positioned to serve as a dedicated venue for this emerging field, supporting research that advances evaluation methodologies, benchmarking, and real-world deployment of AI and language models.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017, Vol. 30, pp. 5998–6008. https://doi.org/10.5555/3295222.3295349.
2. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. Technical report, 2018. Accessed: 2026-03-17.
3. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. Language models are unsupervised multitask learners. Technical report, 2019. Accessed: 2026-03-17.
4. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems, 2020, Vol. 33, pp. 1877–1901. https://doi.org/10.5555/3495724.3495883.
5. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc., 2020, Vol. 33, pp. 9459–9474. Page range from common NeurIPS citation; verify against official proceedings metadata (to be completed).
6. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; et al. Training Language Models to Follow Instructions with Human Feedback. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35.
7. OpenAI. Introducing ChatGPT. OpenAI product blog post, 2022. Accessed: 2026-03-17.
8. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
9. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* **2023**.
10. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. Qwen technical report. *arXiv preprint arXiv:2309.16609* **2023**.
11. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* **2023**, *36*, 46595–46623.
12. Huang, Q.; Tao, M.; Zhang, C.; An, Z.; Jiang, C.; Chen, Z.; Wu, Z.; Feng, Y. Lawyer LLaMA Technical Report, 2023, [arXiv:cs.CL/2305.15062].
13. Li, M.; Zhao, Y.; Yu, B.; Song, F.; Li, H.; Yu, H.; Li, Z.; Huang, F.; Li, Y. API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Bouamor, H.; Pino, J.; Bali, K., Eds., Singapore, 2023; pp. 3102–3116. https://doi.org/10.18653/v1/2023.emnlp-main.187.
14. Huang, Q.; Vora, J.; Liang, P.; Leskovec, J. Benchmarking large language models as ai research agents. In Proceedings of the NeurIPS 2023 foundation models for decision making workshop, 2023.
15. Chan, C.M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; Liu, Z. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* **2023**.
16. Nasution, A.H.; Onan, A. ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access* **2024**, *12*, 71876–71900. https://doi.org/10.1109/ACCESS.2024.3402809.
17. Bonn, J.; Madabushi, H.T.; Hwang, J.D.; Bonial, C. Adjudicating LLMs as PropBank Annotators. 2024, p. 112 – 123. Cited by: 4.
18. He, X.; Lin, Z.; Gong, Y.; Jin, A.L.; Zhang, H.; Lin, C.; Jiao, J.; Yiu, S.M.; Duan, N.; Chen, W. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. 2024, Vol. 6, p. 165 – 190. Cited by: 46; All Open Access, Gold Open Access, https://doi.org/10.18653/v1/2024.naacl-industry.15.
19. Liang, C.; Yan, S.; Guo, Q.; Sun, J.; Cheng, F.; Liu, Z.; Xiao, W.; Wang, M.; Wang, H.; Zhao, X. AEC-Bench: A Multidisciplinary Multilevel Chinese Evaluation Benchmark for Large Language Models in AEC. 2024, p. 137 – 146. Cited by: 0, https://doi.org/10.1061/9780784486115.014.
20. Singh, A.; Fry, A.; Perelman, A.; Tart, A.; Ganesh, A.; El-Kishky, A.; McLaughlin, A.; Low, A.; Ostrow, A.; Ananthram, A.; et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267* **2025**.
21. Crupi, G.; Tufano, R.; Velasco, A.; Mastropaolo, A.; Poshyvanyk, D.; Bavota, G. On the Effectiveness of LLM-as-a-Judge for Code Generation and Summarization. *IEEE Transactions on Software Engineering* **2025**, *51*, 2329 – 2345. Cited by: 3, https://doi.org/10.1109/TSE.2025.3586082.

22. Onan, A.; Nasution, A.H.; Celikten, T. Toward Reliable Annotation in Low-Resource NLP: A Mixture of Agents Framework and Multi-LLM Benchmarking. *IEEE Access* **2025**, *13*, 211620–211644.
23. Khalila, Z.; Nasution, A.H.; Monika, W.; Onan, A.; Murakami, Y.; Radi, Y.B.I.; Osmani, N.M. Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. *International Journal of Advanced Computer Science and Applications* **2025**, *16*, 1361–1371. If citing the arXiv version as well: arXiv:2503.16581., https://doi.org/10.14569/IJACSA.2025.01602134.
24. Nasution, A.H.; Monika, W.; Onan, A.; Murakami, Y. Benchmarking 21 Open-Source Large Language Models for Phishing Link Detection with Prompt Engineering. *Information* **2025**, *16*, 366. https://doi.org/10.3390/info16050366.
25. Nasution, A.H.; Onan, A.; Murakami, Y.; Monika, W.; Hanafiah, A. Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets. *IEEE Access* **2025**, *13*, 94009–94025. https://doi.org/10.1109/ACCESS.2025.3574629.
26. Zhao, Z.; Wang, D. Evaluation of large language models for the intangible cultural heritage domain. *npj Heritage Science* **2025**, *13*. Cited by: 2; All Open Access, Hybrid Gold Open Access, https://doi.org/10.1038/s40494-025-02013-1.
27. Zhang, B.; Takeuchi, M.; Kawahara, R.; Asthana, S.; Maruf Hossain, M.; Ren, G.J.; Soule, K.; Mai, Y.; Zhu, Y. Evaluating Large Language Models with Enterprise Benchmarks. 2025, Vol. 3, p. 485 – 505. Cited by: 3; All Open Access, Gold Open Access, https://doi.org/10.18653/v1/2025.naacl-industry.40.
28. Shalawati, S.; Nasution, A.H.; Monika, W.; Derin, T.; Onan, A.; Murakami, Y. Beyond BLEU: GPT–5, Human Judgment, and Classroom Validation for Multidimensional Machine Translation Evaluation. *Digital* **2026**, *6*, 8. https://doi.org/10.3390/digital6010008.
29. Onan, A.; Nasution, A.H.; Celikten, T.; Cetin, P. Comparing ChatGPT, Gemini, and Emerging LLMs in Low-Resource Educational Settings: Reasoning Quality, Consistency, and Explainability. *IEEE Access* **2026**, *14*, 32807–32838. https://doi.org/10.1109/ACCESS.2026.3669036.
30. Zhu, J.; Li, J.; Zhao, S.; Deng, Y.; Miao, Y.; Xu, J. Adapting LLMs for biomedical natural language processing: a comprehensive benchmark study on fine-tuning methods. *Journal of Supercomputing* **2026**, *82*. Cited by: 0, https://doi.org/10.1007/s11227-025-08182-x.
31. Karpov, I.; Kirillovich, A.; Goncharova, E.; Parinov, A.; Chernyavskiy, A.; Ilvovsky, D.; Semenova, N.; Sosedka, A.; Lisitsyna, E.; Belkin, M. SynEL: A synthetic benchmark for entity linking. *PLOS ONE* **2026**, *21*. Cited by: 0; All Open Access, Gold Open Access, Green Open Access, https://doi.org/10.1371/journal.pone.0339468.
32. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.t. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020, pp. 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550.
33. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Computing Surveys* **2025**. If citing the journal version, confirm issue/pages (to be completed)., https://doi.org/10.1145/3703155.
34. Snow, R.; O'Connor, B.; Jurafsky, D.; Ng, A. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of the Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008, pp. 254–263.
35. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023. https://doi.org/10.18653/v1/2023.emnlp-main.153.
36. Alabdulwahab, A.; Japic, C.; Le, C.; et al. Comparative Study of Large Language Model Evaluation Frameworks with a Focus on NLP vs LLM-As-A-Judge Metrics. In Proceedings of the IEEE SIEDS, 2025.
37. Wang, Y.; Yuan, J.; Chuang, Y.N.; et al. DHP Benchmark: Are LLMs Good NLG Evaluators? In Proceedings of the Findings of NAACL, 2025.
38. Crupi, G.; Tufano, R.; Velasco, A.; et al. On the Effectiveness of LLM-as-a-Judge for Code Generation and Summarization. *IEEE Transactions on Software Engineering* **2025**.
39. Hada, R.; Gumma, V.; de Wynter, A.; et al. Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? In Proceedings of the Findings of EACL, 2024.
40. Li, J.; Sun, S.; Yuan, W.; et al. Generative Judge for Evaluating Alignment. In Proceedings of the ICLR, 2024.
41. Siri, D.; Anuragini, S.; Babu, B.S.; et al. Investigating Benchmarking Techniques for Unveiling Large Language Model Performance. In Proceedings of the AIP Conference Proceedings, 2025.