The OSCAR PUBLISHING Services
American Journal of Applied Science and Technology

VOLUME Vol.05 Issue 11 2025
PAGE NO. 137-143

Integrated Framework For Robust, Responsible, And Evaluative Deployment Of Large Language Models: A Comprehensive Methodological And Theoretical Treatise

Dr. Arjun Sen
Institute for Computational Linguistics, New Delhi University, India

Received: 05 October 2025; **Accepted:** 03 November 2025; **Published:** 27 November 2025

Abstract: Background: The rapid evolution of large language models (LLMs) has outpaced standardized frameworks for their evaluation, deployment, and governance. Diverse evaluation protocols, emergent alignment techniques, and domain adaptation strategies have been proposed, yet a unified, theoretically grounded, and practically applicable framework that connects evaluation, fine-tuning, bias assessment, and end-to-end testing remains underdeveloped.
Objective: This article proposes and elaborates an integrated framework that synthesizes state-of-the-art evaluation methodologies, bias and truthfulness assessments, domain-specific fine-tuning practices, and automation frameworks for end-to-end testing, grounded in existing literature and empirical benchmarks.
Methods: Drawing strictly on the provided literature, we construct a conceptual pipeline where standardized evaluation metrics (including human-aligned LLM-based evaluators), bias and safety assays, and domain adaptation workflows interlock to produce responsible deployment. We analytically extend evaluation taxonomies, compare model families (closed versus open foundation models), and propose best-practice procedural guidelines for automated testing.
Results: The framework clarifies relationships among intrinsic metrics (e.g., perplexity and proxy linguistic measures), extrinsic metrics (task performance), human-aligned automated evaluation (G-Eval and OmniEvalKit principles), and qualitative safety/bias tests (StereoSet, CrowS-Pairs, TruthfulQA). We articulate methodological choices for domain fine-tuning and provide an automation blueprint for continuous validation and regression testing in production settings.
Conclusions: The proposed integrated framework operationalizes evaluation and governance for LLMs, balancing performance optimization with societal risk mitigation. Adoption of this framework can reduce deployment failures, improve alignment with human judgments, and create a replicable pipeline for domain-specialized LLMs. Limitations include dependence on evolving evaluation tools and the need for empirical calibration in diverse application domains. The article closes with prioritized avenues for future research, including benchmark harmonization and adaptive testing regimes.

Keywords: Large language models; evaluation framework; bias assessment; automated testing; human-aligned evaluation; domain fine-tuning.

INTRODUCTION:

The last half-decade has witnessed exponential growth in the capabilities and uptake of large language models (LLMs). From early masked-language pretraining approaches to the emergence of foundation models and chat-oriented generative systems, the research community and industry have prioritized both the expansion of scale and the refinement of architectures (Touvron et al., 2023; Achiam et al., 2023). This expansion has generated novel opportunities and complex challenges. On the opportunity side, LLMs enable few-shot instruction following, emergent reasoning, and wide-ranging downstream applications; on the challenge side, they present pressing concerns about evaluation fidelity, alignment with human preferences, social biases,

truthfulness, and operational reliability (Chang et al., 2024; Srivastava et al., 2022).

The problem statement motivating this work is straightforward but multifaceted: current practices for assessing, fine-tuning, and operationally validating LLMs are fragmented, often inconsistent, and sometimes misaligned with human expectations. Research communities have produced a plethora of targeted evaluation datasets and metrics—each illuminating important facets of model behavior (Lin, Hilton, & Evans, 2021; Nadeem, Bethke, & Reddy, 2020; Nangia et al., 2020; Zellers et al., 2019). Simultaneously, automated evaluators that leverage strong LLMs as judges (e.g., G-Eval) have been proposed to scale assessment while approximating human judgment (Liu et al., 2023). Toolsets like OmniEvalKit aim to modularize evaluation across tasks and modalities (Zhang et al., 2024). Parallel to these developments, practical deployment requires domain-specific adaptation and rigorous end-to-end testing frameworks that can continuously evaluate both performance and safety properties (Bhatnagar, 2023; Chandra, Lulla, & Sirigiri, 2025).

A literature gap persists: there is no holistic, theoretically informed framework that synthesizes these strands—evaluation metrics, alignment techniques, bias measurement, domain fine-tuning, and automation of testing—into an operational pipeline that can be adopted by researchers and practitioners alike. This gap leads to several recurring problems: inconsistent evaluation choices that render cross-study comparisons difficult (Chang et al., 2024), over-reliance on narrow metrics that do not capture human-aligned behavior (Liu et al., 2023), insufficient bias testing in real-world contexts (Nadeem et al., 2020; Nangia et al., 2020), and ad hoc operational testing that fails to detect regression or safety violations in production (Chandra et al., 2025).

This article aims to fill that gap by proposing an integrated framework grounded in extant evaluation research and practical deployment studies. The framework synthesizes theoretical considerations with operational guidance: it prescribes a taxonomy of evaluative dimensions, recommends an evaluation ladder comprising both human and automated judges, articulates bias and truthfulness assays, prescribes domain adaptation workflows for responsible fine-tuning, and outlines an automation framework for continuous testing and monitoring (Liu et al., 2023; Zhang et al., 2024; Bhatnagar, 2023; Chandra et al., 2025). The synthesis draws on foundational model analyses (Achiam et al., 2023; Touvron et al., 2023), examinations of capability extrapolation (Srivastava et al., 2022), and established bias and truthfulness challenge sets (Lin et al., 2021; Nadeem et al., 2020; Nangia et al., 2020; Zellers et al., 2019).

In the remainder of this paper, we first describe the methodological principles used to synthesize the framework. We then present the framework itself, offering detailed explanations of each component and how different evaluation tools interoperate. We follow with a descriptive results section that explains how the components relate to empirical behaviors reported in the literature and what operational outcomes to expect. We conclude with a discussion that interprets the framework, outlines limitations, and suggests future research agendas.

METHODOLOGY

The methodology adopted here is conceptual synthesis and prescriptive framework design grounded strictly in the supplied literature. The aim is not to run new experiments but to systematically analyze and integrate findings, tools, and best practices from the provided references to create an operational, theoretically coherent framework for evaluation and deployment of LLMs. The strategy has three main stages: (1) taxonomic analysis of evaluative constructs; (2) synthesis of evaluative instruments and their trade-offs; (3) procedural design for domain fine-tuning and automation testing.

Taxonomic Analysis of Evaluative Constructs. We first classify evaluation objectives into orthogonal but interconnected dimensions. Drawing on Chang et al. (2024), Srivastava et al. (2022), and Lin et al. (2021), we separate evaluation into intrinsic linguistic competence, task-specific performance, alignment with human values and truthfulness, and societal risk attributes such as bias and toxicity. Intrinsic competence encompasses measures that reflect a model's language modeling capacity and internal representations; task-specific performance refers to downstream metrics for classification, generation, or decision tasks; alignment/truthfulness captures the degree to which outputs correspond to factual reality and human norms; societal risk covers the propensity to generate stereotypical, discriminatory, or harmful content.

Synthesis of Evaluative Instruments and Trade-offs. Informed by the instrumentation literature—ranging from classical metrics like METEOR (Banerjee & Lavie, 2005) to contemporary LLM-based evaluators (Liu et al., 2023) and toolkits (Zhang et al., 2024)—we map each evaluative instrument onto the taxonomy above. This mapping highlights strengths and failure modes—for instance, that lexical overlap metrics

often poorly correlate with human preference in open-ended generation, while LLM-based evaluators scale but can inherit biases from the underlying judge model (Liu et al., 2023; Chang et al., 2024).

Procedural Design for Domain Fine-Tuning and Automation Testing. The final stage constructs procedural workflows for (a) domain-specific fine-tuning and calibration (Bhatnagar, 2023; B. Anil et al., Palm 2 Technical Report, 2023), and (b) continuous automated testing and regression frameworks that operationalize the verification of performance and safety properties (Chandra et al., 2025). We synthesize data curation and prompt design recommendations, exemplar selection for few-shot learning, and validation regimes that combine unit tests, scenario tests, and adversarial stress tests.

Ethical Constraints and Citation Discipline. Because the brief requires strict reliance on the supplied references, all claims and procedural recommendations are referenced to the corresponding literature. Every major claim in the conceptual design is therefore anchored to cited prior work, ensuring traceability and intellectual integrity.

Framework Design Principles. The framework adheres to several design principles synthesized from the literature: modularity (evaluate components independently and jointly; Zhang et al., 2024), multi-perspective assessment (combine humans, automated LLM judges, and standardized benchmarks; Liu et al., 2023), progressive validation (from intrinsic to extrinsic measures; Chang et al., 2024), bias-first testing (prioritize social bias assays early in validation cycles; Nadeem et al., 2020; Nangia et al., 2020), and automation for scale while preserving human-in-the-loop checks for high-stakes outcomes (Chandra et al., 2025).

RESULTS

The "results" here are descriptive outputs of the synthesis process: concrete mappings, procedural checklists, and expectations for model behavior when following the integrated framework. We present the framework in modular subsections and then describe anticipated impacts on evaluation fidelity, alignment, and deployment risk.

A. Evaluative Taxonomy and Instrument Mapping

1. Intrinsic Linguistic Competence. Measures in this class include perplexity-based diagnostics, contextual embedding analyses, and masked language modeling probes (Touvron et al., 2023; Srivastava et al., 2022). These diagnostics are valuable for early-stage model selection and pre-deployment sanity checks. They reveal general capacity but do not guarantee task performance or truthfulness. When paired with probing methods, they enable researchers to diagnose representation gaps that correlate with downstream failures.

2. Task-Specific Performance. Standardized datasets and task benchmarks—ranging from multiple-choice reasoning tests (HellaSwag; Zellers et al., 2019) to domain-specific evaluation suites—fall here. Task-specific metrics provide direct utility measures for a target application but can be gamed by overfitting or prompt engineering. The literature emphasizes careful validation splits and cross-evaluation to prevent misinterpretation of improved benchmark performance as universal capability (Srivastava et al., 2022).

3. Human-Aligned Automated Evaluation. Tools like G-Eval (Liu et al., 2023) operationalize the use of strong LLMs (e.g., GPT-4 families described in Achiam et al., 2023) as automated judges. Advantages include scalability and a better approximation of human preference compared to pure lexical overlap metrics (Banerjee & Lavie, 2005). Risks include judge-model biases and over-reliance on a single judge archetype, potentially propagating systematic errors. Mitigation strategies include ensembling multiple judge models and calibrating against human annotations.

4. Bias, Fairness, and Truthfulness Assays. Datasets such as StereoSet (Nadeem et al., 2020), CrowS-Pairs (Nangia et al., 2020), and TruthfulQA (Lin et al., 2021) provide operational tasks to measure social biases and tendencies to reproduce human falsehoods. These assays are essential pre-deployment, especially for applications with public-facing outputs. They must be interpreted as stress tests rather than definitive certifications—models may pass some bias tests yet fail in domain-specific cultural contexts (Nadeem et al., 2020; Nangia et al., 2020).

B. Procedural Integration and Pipeline Steps

We propose a pipeline with explicit stages, each with objectives and recommended instruments:

1. Pretraining Assessment Phase. Objective: determine whether a base model's internal representations are sufficient for the intended application. Instruments: intrinsic competence diagnostics and probing; cross-reference with foundational model reports (Touvron et al., 2023; Achiam et al., 2023).

2. Controlled Fine-Tuning Phase. Objective: adapt the base model to domain data while mitigating catastrophic forgetting and preserving generalization. Instruments: domain-adaptive fine-tuning workflows

(Bhatnagar, 2023; Palm 2 Technical Report, 2023), held-out validation on both domain tasks and general benchmarks.

3. Human-Aligned Evaluation Phase. Objective: measure alignment to human preference and functional quality. Instruments: G-Eval-like automatic judges for scale, complemented by curated human annotation in critical cohorts (Liu et al., 2023).

4. Bias and Safety Stress-Testing Phase. Objective: proactively detect stereotypical, toxic, or untruthful outputs under adversarial prompts and real-world prompts. Instruments: StereoSet, CrowS-Pairs, TruthfulQA, and adversarial generation techniques (Nadeem et al., 2020; Nangia et al., 2020; Lin et al., 2021).

5. End-to-End Automation and Regression Testing Phase. Objective: deploy an automated testing harness that runs unit-like assertions, scenario tests, and regression checks for performance and safety on continuous delivery. Instruments and principles: automation frameworks and test harnesses described by Chandra et al. (2025) and typical software testing discipline (Bhatnagar, 2023). This layer must integrate monitoring signals for drift and human-in-the-loop escalation for flagged failures.

C. Automation Framework Design

Chandra et al. (2025) outline automation practices for end-to-end testing of LLMs. Building from that, the proposed automation blueprint includes:

1. Test Suite Composition. Combine unit tests (e.g., deterministic mapping examples), scenario tests (realistic user interactions), and stochastic robustness tests that sample generation distributions.

2. Continuous Integration Hooks. Integrate test runs into model deployment pipelines so that every model version triggers a full battery of tests.

3. Alerting and Rollback Policies. Define explicit thresholds for failing test categories (e.g., bias score increases beyond a limit) to automatically block deployment and notify human reviewers.

4. Model Evaluation Dashboarding. Provide real-time dashboards for metric trends and anomaly detection, enabling proactive governance decisions.

D. Interaction Effects and Trade-offs

Following the literature, we describe trade-offs practitioners must navigate:

1. Performance versus Safety. Aggressive fine-tuning to optimize for a narrow metric can increase propensity for overconfident or unsafe outputs (Srivastava et al., 2022). A balanced objective function that incorporates safety-aware losses or post-hoc filtering is necessary.

2. Scale and Interpretability. Larger models often gain capabilities but become harder to inspect and calibrate (Touvron et al., 2023). The framework recommends complementary probe analyses and modular evaluation to preserve interpretability.

3. Automated Judges and Inherited Bias. Leveraging strong LLMs as judges scales human-alignment evaluation but risks inheriting their biases and blind spots (Liu et al., 2023). Ensemble judging and periodic human calibration are recommended to reduce risk.

E. Expected Outcomes from Framework Adoption

Adopting this integrated framework should yield several operational outcomes:

1. Harmonized Evaluation: Better comparability across studies through multi-dimensional reports that include intrinsic, extrinsic, alignment, and safety metrics (Chang et al., 2024; Zhang et al., 2024).

2. Reduced Deployment Failures: Automated testing and stricter regression policies reduce catastrophic deployment errors and social harm incidents (Chandra et al., 2025).

3. Human-Preferred Outputs: Using LLM-based judges calibrated to human annotations can improve the alignment of deployed outputs with user expectations (Liu et al., 2023).

4. Faster Domain Adaptation: Protocolized fine-tuning workflows accelerate domain-specific deployment while preserving generalization (Bhatnagar, 2023; Palm 2 Technical Report, 2023).

DISCUSSION

This section interprets the framework more deeply, addresses limitations, positions the contribution relative to existing literature, and articulates a research and operational agenda. The primary contribution of this work is synthesis: integrating disparate evaluation tools, bias assays, domain adaptation practices, and automation frameworks into a single, actionable pipeline. We discuss each aspect's theoretical implications and potential counter-arguments.

Theoretical Implications

1. Multi-dimensional Evaluation as Epistemic Guardrail. The taxonomy proposed—spanning intrinsic competence, task performance, human-aligned evaluation, and societal risk—functions as an epistemic guardrail. By requiring evidence across orthogonal measures, the framework mitigates the epistemic overconfidence that arises when a model

excels on one measure but fails in other dimensions (Chang et al., 2024; Srivastava et al., 2022). This aligns with the philosophical principle that robustness requires multiple independent lines of evidence.

2. Human-Aligned LLM Judges as a Pragmatic Intermediate. The use of strong LLMs (e.g., in G-Eval) as automated judges occupies a middle ground between expensive human annotation and brittle lexical metrics. Theoretically, this introduces a second-order modeling phenomenon: we evaluate a model using another model that embeds human-like preferences. This has two implications. First, it scales evaluation and can better correlate with human judgment (Liu et al., 2023). Second, it risks propagating systemic biases present in judge models, necessitating calibration and ensembling as corrective mechanisms.

3. Bias Assays as Necessary but Insufficient Diagnostics. Tools like StereoSet and CrowS-Pairs are indispensable for probing stereotypical associations, but they are constructed around specific cultural frames and tasks (Nadeem et al., 2020; Nangia et al., 2020). The framework therefore treats such assays as early detection tools that must be supplemented by domain- and culture-specific checks. Theoretical rigor demands that bias detection not be reduced to numeric thresholds alone; contextual human review is required to interpret tests' real-world implications.

4. Continuous Testing as Socio-Technical Practice. The automation framework reframes testing not as a one-off pre-deployment activity but as a socio-technical practice that combines technical monitoring with governance processes (Chandra et al., 2025). This perspective aligns with contemporary thinking in software engineering—systems must be continuously validated in production where distributional shifts occur.

Counter-Arguments and Nuances

Several counter-arguments meriting attention follow from the literature. One could argue that reliance on LLM-based judges simply replaces human fallibility with model fallibility. This critique is valid and recognized within the framework; thus, we recommend hybridization—automated judges for scale, but with targeted human audits in high-stakes or ambiguous domains (Liu et al., 2023). Another counterpoint contends that standardized benchmarks incentivize narrow optimization and benchmark gaming (Srivastava et al., 2022). The framework addresses this by recommending diverse benchmark suites, cross-dataset generalization checks, and adversarial testing to detect overfitting to benchmarks.

Practical Limitations

No framework is a panacea. We highlight several realistic limitations.

1. Evolving Evaluation Tools. The field's tools evolve rapidly (Zhang et al., 2024; Chang et al., 2024). The framework must therefore be treated as a living document; specific instruments cited here (e.g., G-Eval, OmniEvalKit) will require periodic re-evaluation and substitution as the state of the art advances.

2. Resource Constraints. Rigorous testing, including extensive human annotation and continuous integration, entails costs that may be infeasible for smaller teams or low-resource contexts. The framework's modularity is thus critical: practitioners should prioritize core checks (bias, truthfulness, human-aligned evaluation) and gradually scale up.

3. Cultural and Domain Specificity. Tests developed in one cultural or linguistic context may not generalize. The framework mandates local adaptation and cultural sensitivity in bias and safety testing (Nadeem et al., 2020; Nangia et al., 2020).

4. Judge Calibration. Using LLMs as judges presupposes that judge models are themselves aligned and high-quality (Achiam et al., 2023). Without access to such judges or when judge model biases are unknown, calibration becomes difficult.

Bridging the Gap to Practice

To facilitate adoption, the framework recommends concrete policies:

1. Evaluation Contracts. Before development, teams should define an "evaluation contract" specifying required tests, acceptable thresholds, and remediation plans for failures. This contract formalizes expectations and governance.

2. Release Notes and Metric Transparency. Model releases must accompany exhaustive metric reports covering all taxonomic dimensions, including raw distributions and failure cases (Chang et al., 2024).

3. Human Oversight Protocols. For high-stakes outputs, automatic blocking mechanisms should trigger human review. The automation framework must specify who reviews, criteria for escalation, and timelines (Chandra et al., 2025).

4. Post-Deployment Monitoring. Continuous monitoring for drift, user feedback, and emergent bias is crucial. Metrics should be designed to capture subtle shifts in behavior after deployment.

Future Research Directions

The literature suggests several promising trajectories:

1. Benchmark Harmonization. The proliferation of benchmarks complicates comparison. Future work should focus on aligning benchmark taxonomies, constructing meta-benchmarks, and establishing cross-walks between datasets (Chang et al., 2024; Zhang et al., 2024).

2. Judge Model Governance. The community must study the properties of judge models and their biases comprehensively to improve the reliability of automated evaluation (Liu et al., 2023).

3. Cultural Sensitivity in Bias Testing. More effort is required to build bias and safety tests that are culturally nuanced and linguistically inclusive (Nadeem et al., 2020; Nangia et al., 2020).

4. Theory of Model Alignment. Theoretical work is needed to formalize the relationship between pretraining distributions, fine-tuning regimes, and alignment outcomes (Srivastava et al., 2022; Achiam et al., 2023).

5. Cost-Aware Testing Strategies. Develop testing strategies that achieve acceptable safety and performance levels under constrained budgets.

CONCLUSION

This article presents an integrated, theoretically grounded, and operationally pragmatic framework for evaluating, fine-tuning, and deploying large language models. By synthesizing contemporary developments in evaluation methodologies, automated LLM-based judging, bias-assessment datasets, domain fine-tuning practices, and automation frameworks for testing, the framework offers a replicable pipeline for responsible LLM deployment. Key recommendations include multi-dimensional evaluation, hybrid human-automated judging calibrated to human annotations, early bias and truthfulness stress-testing, protocolized domain adaptation workflows, and robust automation for continuous validation and regression testing. The framework's adoption promises harmonized evaluation reporting, reduced deployment failures, and outputs better aligned with human preferences and societal norms. Limitations include dependence on evolving tools, resource constraints, and cultural specificity; mitigation strategies emphasize modularity and human oversight. Future research should pursue benchmark harmonization, judge model governance, culturally sensitive bias testing, theoretical models of alignment, and cost-aware testing approaches. This synthesis aims to serve as a decision-oriented roadmap linking theoretical insights and practical engineering disciplines to foster safer, more reliable, and human-centered LLM systems.

REFERENCES

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.
2. Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). PaLM 2 technical report.
3. Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65–72).
4. Bhatnagar, D. (2023). Fine-Tuning Large Language Models for Domain-Specific Response Generation: A Case Study on Enhancing Peer Learning in Human Resource. PhD thesis, Dublin, National College of Ireland.
5. Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. Education and Information Technologies, 30(2):2041–2058.
6. Chandra, R., Lulla, K., & Sirigiri, K. (2025). Automation frameworks for end-to-end testing of large language models (LLMs). Journal of Information Systems Engineering and Management, 10, e464-e472.
7. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), 1–45.
8. Chen, Z., Balan, M. M., & Brown, K. (2023). Language models are few-shot learners for prognostic prediction. arXiv preprint arXiv:2302.12692.
9. Crain, P., Lee, J., Yen, Y.-C., Kim, J., Aiello, A., & Bailey, B. (2023). Visualizing topics and opinions helps students interpret large collections of peer feedback for creative projects. ACM Transactions on Computer-Human Interaction, 30(3).
10. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. arXiv preprint arXiv:2303.16634.
11. Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958.
12. Nadeem, M., Bethke, A., & Reddy, S. (2020).

StereoSet: Measuring stereotypical bias in pretrained language models. arXiv preprint arXiv:2004.09456.

13. Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. arXiv preprint arXiv:2010.00133.
14. PaLM 2 Technical Report (2023). R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al.
15. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Wang, G. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.
16. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
17. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830.
18. Zhang, Y. K., Zhong, X. X., Lu, S., Chen, Q. G., Zhan, D. C., & Ye, H. J. (2024). OmniEvalKit: A modular, lightweight toolbox for evaluating large language models and its omni-extensions. arXiv preprint arXiv:2412.06693.