

Uncertainty-Aware Reasoning Path Filtering for Reliable Large Language Model Inference

Aleks Tomalik

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
aleks820@uc.edu

Elsie Simonsen

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
contactelsie@missouri.edu

Almira Widodo

School of Computing, Clemson University, Clemson, SC, USA.
almirawork@clemson.edu

Abstract

Large language models increasingly operate as general-purpose reasoning engines in high-stakes environments, yet their outputs remain difficult to audit, calibrate, and govern. This paper examines uncertainty-aware reasoning path filtering as a system-level strategy for improving the reliability of large language model inference. Rather than treating uncertainty as a post hoc confidence score attached to a final answer, the discussion positions uncertainty as a structural signal distributed across multiple candidate reasoning paths. The paper develops an architectural account of reasoning path generation, scoring, filtering, and escalation, with attention to the trade-offs that emerge among accuracy, latency, computational cost, fairness, and operational sustainability. It argues that reliable inference depends less on a single model improvement than on the surrounding socio-technical infrastructure that decides when to trust, revise, defer, or reject generated reasoning. The analysis connects calibration, out-of-distribution detection, ensemble disagreement, and verbalized uncertainty to practical deployment concerns such as auditability, model documentation, human oversight, and environmental cost. It also considers how filtering mechanisms may amplify or mitigate inequities when confidence estimates are unevenly distributed across languages, domains, and populations. The paper concludes that uncertainty-aware reasoning path filtering offers a promising but incomplete foundation for reliable inference, and that its benefits depend on transparent governance, cross-domain validation, and infrastructure capable of supporting adaptive computation without creating new forms of exclusion or hidden technical debt.

Keywords

uncertainty-aware inference; reasoning path filtering; large language models; reliable inference; system architecture; calibration; AI governance; deployment sustainability.

1. Introduction

Large language models have moved from experimental text generators to infrastructural components of search, software development, scientific assistance, customer service, education, and public administration. The transformer architecture provided the scalable

attention mechanism that made this transition technically feasible, while large-scale pretraining demonstrated that broad linguistic and task competence could emerge without task-specific supervision [1], [2]. More recent reasoning-oriented prompting methods have extended these capabilities by encouraging models to produce intermediate steps, compare alternatives, and revise initial answers before committing to a final response [3], [4]. These developments have created a new reliability problem: a model may appear fluent and coherent while following a reasoning path that is unstable, poorly calibrated, or unsafe in context.

The central concern of this paper is not whether a single generated answer is correct, but how an inference system can decide which reasoning trajectories deserve trust. Uncertainty-aware reasoning path filtering treats generated reasoning as a population of candidate paths rather than a single opaque output. Paths are generated, evaluated, compared, and selectively retained according to signals such as model confidence, self-consistency, ensemble disagreement, verbalized hesitation, and distributional novelty. This approach is system-level because it requires orchestration among model components, scoring mechanisms, fallback policies, human review channels, and monitoring infrastructure. It also raises structural trade-offs: more filtering can improve reliability but increase latency and cost, while aggressive early stopping can reduce computational burden but suppress useful deliberation.

The paper argues that uncertainty-aware reasoning path filtering should be understood as a socio-technical infrastructure rather than a narrow algorithmic patch. Reliable inference depends on how uncertainty is represented, where it is computed, who interprets it, and what institutional actions follow from it. A model that reports low confidence but has no pathway to escalation may be no safer than a model that reports false certainty. Conversely, a system that escalates every ambiguous case to human review may be reliable in principle but unsustainable in practice. The following sections develop this argument by examining uncertainty as a system property, the architecture of reasoning paths, filtering mechanisms, deployment trade-offs, fairness and governance implications, and cross-domain lessons for future research.

2. Uncertainty as a System Property in Large Language Model Inference

Uncertainty in large language model inference is often discussed as if it were a single scalar attached to a prediction. In practice, it is a multi-layered property of the entire inference pipeline. Data uncertainty reflects ambiguity inherent in the input, such as an underspecified question or conflicting evidence. Model uncertainty reflects limitations in parameters, training data, or architectural capacity. Distributional uncertainty reflects mismatch between the deployment context and the conditions under which the model was trained or evaluated. System uncertainty reflects latency, retrieval failures, tool errors, or unstable orchestration logic. A reasoning path may therefore be uncertain for different reasons, and the appropriate response depends on which layer is responsible.

Calibration research provides one foundation for understanding these distinctions. Modern neural networks can produce probability estimates that are poorly aligned with empirical correctness, especially under distribution shift [11]. Bayesian approximations such as dropout-based inference and deep ensembles offer ways to estimate predictive uncertainty by measuring variability across model configurations or stochastic passes [12], [13]. Out-of-distribution detection methods further help identify inputs that fall outside the model’s reliable operating region [14]. These techniques are valuable, but they were largely developed for classification and vision tasks. Applying them to open-ended language generation requires

translating uncertainty about tokens, sequences, or semantic claims into uncertainty about reasoning quality.

Language models can also be prompted to express uncertainty in natural language. Studies suggest that models sometimes “know what they know,” meaning that their internal representations or verbalized statements contain information about likely correctness [9]. Other work has explored teaching models to express uncertainty in words, producing hedges, caveats, or confidence qualifiers that human users can interpret [10]. However, verbalized uncertainty is not automatically trustworthy. Models may be fluent in the language of hesitation while remaining systematically overconfident in domains where training data are sparse. This makes uncertainty-aware filtering a hybrid problem: it must combine statistical signals, linguistic signals, and contextual policy signals rather than relying on any single indicator.

From a system perspective, uncertainty should be treated as an actionable state rather than a descriptive score. The system must decide whether to accept a reasoning path, request additional computation, retrieve external evidence, invoke a tool, ask a clarifying question, or escalate to a human. These decisions are governance decisions as much as technical ones. They determine who bears the cost of errors, how quickly service is delivered, and which users receive the benefit of deeper deliberation. Framing uncertainty as a system property therefore shifts attention from model-only improvements to the design of decision boundaries, escalation pathways, and accountability structures.

3. Reasoning Paths and the Architecture of Deliberation

Reasoning path filtering presupposes that a model or system can generate multiple paths toward an answer. Chain-of-thought prompting demonstrated that intermediate reasoning steps can improve performance on tasks requiring arithmetic, commonsense, and symbolic manipulation [3]. Self-consistency extended this idea by sampling multiple reasoning paths and selecting the answer that appears most frequently, effectively using agreement among paths as a reliability signal [4]. Tree of Thoughts and Graph of Thoughts generalize the approach further by allowing branching, backtracking, evaluation, and aggregation across structured reasoning states [5], [6]. These methods transform inference from a single forward pass into a search process over possible lines of thought.

Iterative refinement methods add another layer. Self-refine allows a model to critique and revise its own outputs, while Reflexion incorporates verbal feedback and episodic memory to improve future attempts [7], [8]. Such methods can produce better answers, but they also increase the number of model calls and the complexity of orchestration. A system that generates many paths, critiques them, revises them, and then selects among them is no longer a simple model endpoint. It is a computational workflow with scheduling, caching, budget limits, and failure modes. The reliability of the final answer depends on the reliability of each stage, including the evaluator that judges which path is best.

The architecture of deliberation therefore matters as much as the base model. If all reasoning paths are generated by the same model under the same prompt, they may share correlated errors. Diverse sampling temperatures can increase variation, but they can also introduce incoherence. External tools, retrieval systems, or verifier models can provide independent evidence, but they introduce new dependencies and potential points of failure. A well-designed system must balance exploration and exploitation: it needs enough diversity to

detect uncertainty, but enough consistency to produce a stable final response. Reasoning path filtering is the mechanism that mediates this balance.

Moreover, reasoning paths are not merely internal computational artifacts. They can be logged, audited, displayed, or summarized for users. This creates opportunities for transparency but also risks information overload. A user may not benefit from seeing ten competing rationales if the system cannot explain why one was selected. In regulated settings, however, the ability to inspect the path may be essential for accountability. The design challenge is to preserve enough path-level evidence for audit and redress while presenting a final answer that is coherent and actionable.

4. Uncertainty-Aware Filtering as a Structural Mechanism

Uncertainty-aware filtering can be understood as a set of structural decisions about which reasoning paths survive to influence the final output. At the generation stage, the system may sample multiple candidate paths, each with its own intermediate claims and conclusions. At the scoring stage, each path receives signals from self-consistency, model confidence, external verification, or learned reward models. At the filtering stage, paths that fall below a threshold may be discarded, while paths that exceed a threshold may be merged or selected. At the escalation stage, unresolved uncertainty may trigger retrieval, tool use, human review, or abstention.

Recent work on scalable reasoning path filtering for smaller parameter models illustrates the practical appeal of this approach [15]. The core insight is that not every reasoning trajectory deserves equal computational investment. By filtering paths early, a system can reduce waste and align the model toward more reliable patterns without requiring enormous parameter counts. This is particularly relevant for deployment contexts where latency, memory, and energy are constrained. However, the same logic raises concerns about premature convergence. If the filter is too aggressive, it may discard unconventional but correct paths. If it is too permissive, it may preserve noisy reasoning that misleads the final answer.

The filtering mechanism must therefore be calibrated not only to accuracy but also to the cost of different error types. In medical triage, a false negative may be far more dangerous than a false positive, so the filter should favor recall and escalation. In creative ideation, excessive filtering may suppress novelty and diversity. In legal analysis, the system may need to preserve dissenting interpretations for human review. A single global threshold is unlikely to satisfy these varied requirements. Effective uncertainty-aware filtering requires context-sensitive policies that reflect domain risk, user expectations, and regulatory obligations.

A further structural issue is where filtering occurs. Early filtering reduces computation but may rely on weak signals. Late filtering allows richer evaluation but consumes more resources. Cascading architectures can combine both approaches, using cheap heuristics to eliminate obviously poor paths and expensive verification for the remaining candidates. This layered design resembles reliability engineering in other critical systems, where redundancy, monitoring, and fail-safe defaults are distributed across components. The goal is not perfect certainty but graceful degradation: the system should behave predictably when uncertainty is high, rather than silently producing confident but unreliable output.

5. Trade-Offs Among Reliability, Latency, Cost, and Scale

The primary promise of uncertainty-aware reasoning path filtering is improved reliability, but reliability is not free. Generating multiple paths multiplies inference cost. Scoring and

filtering require additional model calls or auxiliary evaluators. Escalation to human review introduces operational expense and delay. In high-volume applications, even small per-query overheads can become substantial. System architects must therefore decide how much deliberation is justified by the value of the task and the cost of an error. A customer service chatbot and a radiology decision support tool may both use large language models, but their acceptable trade-offs are radically different.

Latency is a particularly visible constraint. Users expect interactive systems to respond quickly, yet deep reasoning often requires sequential computation. Filtering can reduce latency by stopping low-quality paths early, but it can also increase latency by adding evaluation steps. One design response is adaptive computation: the system allocates more reasoning effort only when uncertainty is high or when the task is deemed important. This approach can be efficient, but it creates a two-tier experience. Users whose queries are classified as easy may receive fast answers, while users whose queries trigger uncertainty may experience delays or be routed to human agents. If uncertainty correlates with dialect, language, or domain, this tiering may produce inequitable service.

Cost and scale are also entangled with environmental sustainability. Large-scale inference already consumes significant energy and water for cooling, and reasoning-intensive workflows amplify these demands [16]. Filtering can reduce waste by avoiding unnecessary paths, but only if the filtering itself is cheap. If the system uses another large model to evaluate every path, the net savings may be negative. Sustainable design requires careful measurement of the full computational footprint, including generation, scoring, retrieval, logging, and human review. It also requires transparency about when deeper reasoning is truly necessary and when a simpler model or rule-based fallback would suffice.

Finally, scale affects governance. A small pilot deployment can rely on manual review and close monitoring. A global deployment cannot. Automated filtering policies must be encoded, versioned, and audited across regions. Different jurisdictions may impose different requirements for transparency, data protection, and contestability. A filtering threshold that is acceptable in one context may be unlawful or unethical in another. Scalable reliability therefore depends on modular policy layers that can be adapted without retraining the entire model. The technical architecture must support governance diversity rather than assume a single universal standard.

6. Governance, Fairness, and Accountability

Uncertainty-aware filtering intersects directly with fairness. If a model is less confident about certain dialects, languages, or cultural contexts because of uneven training data, then uncertainty-based filtering may systematically deprioritize those users. They may receive more caveats, slower responses, or repeated requests for clarification. Over time, such patterns can reinforce existing inequalities in access to reliable information. Fairness-aware filtering must therefore examine not only overall accuracy but also the distribution of uncertainty, escalation, and abstention across groups. A system that is reliable on average may still be unjust in its error allocation.

Accountability requires documentation and traceability. Model cards and datasheets provide structured ways to record intended use, limitations, evaluation data, and performance characteristics [19], [20]. For reasoning path filtering, documentation must go further. It should describe how paths are generated, what uncertainty signals are used, how thresholds are set, and what happens when uncertainty remains unresolved. It should also record known

failure modes, such as correlated path errors or adversarial manipulation of confidence cues. Without such documentation, auditors cannot determine whether a system’s reliability claims are justified.

Policy frameworks for foundation models emphasize both opportunities and risks, including concentration of power, environmental impact, and misuse [16]. Ethical risk taxonomies further identify harms related to misinformation, discrimination, privacy, and human autonomy [18]. Uncertainty-aware filtering can support governance by enabling abstention and human oversight, but it can also obscure responsibility. If a model produces a harmful answer and the system reports that the path was “low confidence,” who is accountable? The developer, the deployer, the reviewer, or the user? Clear governance requires assigning responsibility for threshold design, escalation decisions, and the consequences of filtered or unfiltered outputs.

The broader critique of large language models as “stochastic parrots” warns against treating fluency as understanding [17]. Uncertainty-aware filtering does not resolve this critique by itself. A model can generate a confident-sounding rationale that is merely a plausible reconstruction of patterns in training data. Filtering may select the most coherent path without verifying its truth. Therefore, governance should not overstate the epistemic authority of filtered outputs. Instead, filtering should be positioned as one component in a larger accountability regime that includes external evidence, domain expertise, user education, and avenues for contestation.

7. Deployment, Infrastructure, and Sustainability

Deploying uncertainty-aware reasoning path filtering requires infrastructure that can support dynamic computation. The system needs mechanisms for parallel path generation, caching of intermediate results, efficient scoring, and policy enforcement. It also needs observability: logs, metrics, traces, and alerts that reveal when uncertainty is rising or when filtering is behaving unexpectedly. This resembles the hidden technical debt of machine learning systems, where data dependencies, configuration entanglement, and feedback loops create long-term maintenance burdens [21]. Reasoning filters add another layer of debt because their behavior may drift as models, prompts, and user populations change.

Operational resilience depends on fallback strategies. When uncertainty is high, the system may retrieve documents, call a calculator, consult a database, or transfer to a human. Each fallback introduces dependencies that must be monitored. A retrieval service outage may cause the system to filter out correct paths because external evidence is unavailable. A human review queue may become a bottleneck during peak demand. Resilient design requires redundancy, graceful degradation, and clear service-level expectations. It also requires simulating failure scenarios before deployment, not after a public incident.

Sustainability concerns are not limited to energy. They include the human labor of annotation, review, and maintenance. Uncertainty-aware filtering often relies on labeled data to train evaluators or to calibrate thresholds. This labeling work can be precarious, underpaid, and psychologically taxing. If the system escalates uncertain cases to human reviewers, those reviewers may face a disproportionate share of ambiguous or distressing content. A socially sustainable deployment must account for worker well-being, fair compensation, and limits on cognitive load. Otherwise, reliability gains for users may be achieved through hidden harms to workers.

Infrastructure choices also affect equity. Centralized inference may offer economies of scale and consistent updates, while edge or on-premises deployment may offer privacy and local control. Uncertainty-aware filtering can be implemented in either architecture, but the trade-offs differ. Centralized systems can aggregate uncertainty signals across many users, potentially improving calibration, but they also raise surveillance and data concentration risks. Local systems may protect privacy but have less data for calibration. A polycentric infrastructure that supports both centralized and local components may be more resilient, provided that governance mechanisms can coordinate across boundaries.

8. Case Illustrations and Cross-Domain Comparisons

In healthcare, uncertainty-aware reasoning path filtering could support clinical decision support by generating multiple differential diagnoses, filtering those that conflict with patient history, and escalating uncertain cases to clinicians. The benefit is not autonomous diagnosis but better preparation for human judgment. The risk is automation bias: clinicians may over-trust a filtered output that appears confident. Effective deployment would require explanations that expose path-level uncertainty, clear indications of when the system is outside its competence, and workflows that preserve clinical responsibility.

In public administration, the same approach might be used to assess eligibility for benefits or to draft responses to citizen inquiries. Here, fairness and contestability are paramount. A filtered reasoning path that denies a benefit must be traceable to rules and evidence, not merely to a model's confidence score. Citizens should be able to understand the basis of a decision and challenge it. This suggests that uncertainty-aware filtering in high-stakes administrative settings should favor abstention and human review over automated finalization. The goal is not to maximize automation but to allocate discretion appropriately.

In scientific research, reasoning path filtering could help generate hypotheses, design experiments, or review literature. Uncertainty signals may indicate where evidence is thin or where multiple interpretations are plausible. Filtering can reduce noise, but it may also narrow the hypothesis space prematurely. A system that selects only the most probable path may miss transformative but initially unlikely ideas. Scientific deployment may therefore require diversity-preserving filters that retain minority paths for human consideration. This contrasts with safety-critical domains where conservative filtering is more appropriate.

In education, uncertainty-aware filtering could provide feedback on student reasoning. If the system detects multiple valid solution paths, it can avoid marking a correct but unconventional answer as wrong. If it detects high uncertainty, it can ask the student to explain further rather than issuing a definitive grade. This illustrates a general principle: the value of filtering depends on whether the system's response to uncertainty is pedagogically or procedurally appropriate. Cross-domain comparison shows that there is no universal filtering policy. Reliability is context-dependent and must be designed in relation to domain values.

9. Open Challenges and Research Directions

Several open challenges remain. First, uncertainty signals for language reasoning are not yet well understood. Self-consistency can be misleading when all sampled paths share the same bias. Verbalized confidence can be manipulated by prompt phrasing. Ensemble methods are expensive and may still produce correlated errors. Future research should develop evaluation benchmarks that measure not only answer accuracy but also the quality of uncertainty estimates across diverse reasoning tasks and populations.

Second, filtering introduces a new source of error: the filter itself. If the evaluator is biased or brittle, it may discard correct paths or retain harmful ones. Research is needed on adversarial robustness, distribution shift, and the interaction between filtering and prompt injection. A malicious input might deliberately inflate confidence or suppress uncertainty signals. Defenses must be tested at the system level, not just at the model level.

Third, governance mechanisms for uncertainty-aware inference are immature. Standards for documenting filtering policies, auditing thresholds, and providing contestability are still emerging. Policymakers should consider requirements for transparency about when systems abstain, escalate, or filter. They should also support independent evaluation and red-teaming. Without such governance, uncertainty-aware filtering may become a form of “ethics washing,” giving the appearance of caution while leaving underlying power asymmetries intact.

Fourth, sustainability and equity must be integrated into design from the beginning. It is not enough to measure accuracy gains if those gains come at unacceptable environmental or social cost. Researchers should report computational budgets, human labor impacts, and distributional outcomes alongside performance metrics. Funding agencies and journals can encourage this by requiring system-level reporting. The long-term goal is a reliability practice that is technically robust, institutionally accountable, and socially just.

10. Conclusion

Uncertainty-aware reasoning path filtering offers a promising direction for reliable large language model inference because it treats reliability as a property of the inference system rather than a single output. By generating multiple reasoning paths, evaluating them with uncertainty signals, and selectively filtering or escalating them, systems can better manage ambiguity, avoid premature certainty, and allocate computation where it matters most. This approach aligns with broader advances in chain-of-thought reasoning, self-consistency, iterative refinement, and uncertainty estimation. It also responds to practical deployment needs for auditability, fallback, and human oversight.

However, filtering is not a complete solution. It introduces new trade-offs among accuracy, latency, cost, fairness, and sustainability. It can reproduce existing biases if uncertainty is unevenly distributed across users. It can create hidden technical debt if filtering policies are not documented and maintained. It can obscure accountability if responsibility for escalation and abstention is unclear. The most reliable systems will therefore be those that combine uncertainty-aware filtering with transparent governance, cross-domain validation, worker protections, and meaningful channels for contestation. The future of reliable inference lies not in eliminating uncertainty but in building infrastructures that can recognize it, communicate it, and act on it responsibly.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020).

- Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates.
3. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 24824–24837). Curran Associates.
 4. Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*. OpenReview.
 5. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 11809–11822). Curran Associates.
 6. Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., & Hoefler, T. (2024). Graph of Thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 16, pp. 17682–17690). AAAI Press.
 7. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 46534–46594). Curran Associates.
 8. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 8634–8652). Curran Associates.
 9. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., ... Kaplan, J. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
 10. Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.
 11. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR.
 12. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 1050–1059). PMLR.
 13. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 6402–6413). Curran Associates.

14. Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations. OpenReview.
15. Li, Q. (2026, May). TrajLite PPO: Scalable Reasoning Path Filtering and Alignment for Small Parameter Large Language Models. In 2026 2nd International Conference on Artificial Intelligence and Digital Ethics (ICAIDE) (pp. 29-32). IEEE.
16. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
17. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM.
18. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.
19. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM.
20. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.