Trustworthy Domain-Specific Language Models Through Reinforcement Learning-Based Risk Assessment and Evidence-Aware Response Control

Dex Teunissen
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
teunissen1992@colostate.edu

Ksawery Kasica
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
ksawery.work@uab.edu

Brandy Welch
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
brandy1970@uc.edu

Abstract

Domain-specific language models are increasingly proposed for settings in which an unsupported, misapplied, or overconfident answer can have material consequences. The central problem is not only linguistic quality but calibrated risk assessment and disciplined response control when domain evidence is incomplete, conflicting, or stale. This manuscript develops a methodological framework that couples reinforcement learning with evidence-aware response control for trustworthy domain-specific language models. The framework represents the query, retrieved evidence, provenance, contradiction signals, and interaction history as a decision state; uses a risk-sensitive critic and a calibrated immediate-risk estimator to assess the probability that a candidate response violates domain-specific acceptability conditions; and trains a policy that selects among answer, hedge, cite, retrieve, clarify, abstain, and escalate actions under an explicit risk budget. Evidence sufficiency is modeled separately from answer confidence, so the controller can distinguish an unsupported confident answer from a well-grounded but incomplete one. The formulation includes a calibration objective, a risk-constrained reinforcement learning objective, a conditional value-at-risk constraint, an evidence-aware threshold rule, and a risk-adjusted advantage function. The analytical assessment explains how underestimation of risk, poor evidence calibration, action costs, and reward misspecification affect behavior, and it identifies domain-specific failure modes together with evaluation logic. The contribution is a coherent technical design for response control that treats abstention and escalation as first-class actions rather than post hoc safeguards, with implications for clinical, legal, financial, and other evidence-intensive deployments.

Keywords

risk-sensitive reinforcement learning; retrieval-augmented generation; evidence sufficiency; calibration; abstention; domain-specific language models

1. Introduction

Domain-specific language models (DSLMs) adapt general-purpose language modeling to specialized vocabularies, document genres, and decision contexts. Their appeal lies in combining broad linguistic competence with access to domain corpora, retrieval systems, and task-specific instructions. Instruction tuning and preference-based optimization have improved the usefulness and controllability of such models [1], [2], [3]. These advances, however, do not by themselves establish that a model will answer within the evidential limits of a domain, recognize when retrieved evidence is insufficient, or avoid confident assertions that a domain expert would reject. In high-stakes settings, the relevant question is not whether an answer is fluent but whether its risk of unacceptable error has been assessed before the answer is released.

Retrieval-augmented generation has become a standard mechanism for grounding model output in external text [4], [5], [6]. Retrieval can reduce unsupported generation, but it introduces new failure modes: the retrieved passage may be irrelevant, outdated, contradictory, or applicable only under conditions absent from the query. A model may also overstate support from a weakly related document. Recent work on uncertainty-aware retrieval-augmented generation emphasizes that trustworthy generation requires risk control over the retrieval-and-generation process itself [7]. That observation motivates a tighter coupling between risk estimation and response selection. A system should not merely retrieve and then generate; it should decide whether the available evidence supports an answer, whether the answer should be qualified, whether additional retrieval is needed, or whether the response should be abstained or escalated.

Language models can express uncertainty, and their internal confidence signals carry useful information [8], [9]. Calibration research shows, however, that neural network confidence is often imperfect and can vary markedly across domains and distribution shifts [10]. Hallucination and faithfulness research has documented the persistence of unsupported content even in strong generation systems [11], [12], and fact-verification benchmarks provide a basis for evaluating evidential support [13]. Retrieval-augmented generation surveys further show that retrieval quality, evidence integration, and response verification are distinct design problems [14]. These lines of work establish the ingredients for trustworthy DSLMs but do not provide a unified control formulation in which risk, evidence sufficiency, action cost, and response mode are optimized jointly.

This manuscript develops such a formulation. The central contribution is a reinforcement-learning framework for risk-aware, evidence-conditioned response control. The framework treats a DSLM response as an action in a partially observed decision process whose state includes the query, retrieved evidence, provenance, contradiction signals, and interaction history. A risk-sensitive critic estimates cumulative risk-adjusted value, while a calibrated risk head estimates the immediate probability that a candidate response violates domain-specific acceptability conditions. A separate evidence-sufficiency score captures whether the retrieved evidence covers the factual requirements of the query. The policy is trained to maximize domain utility subject to a risk budget and action costs, with conditional value-at-risk used to constrain the upper tail of the risk distribution. A threshold controller then maps risk and evidence sufficiency into response modes, including answer, hedge, cite, retrieve more, ask for clarification, abstain, and escalate. The analytical development explains why separating evidence sufficiency from answer confidence is necessary, how risk misestimation propagates into response behavior, and how domain-specific error modes should shape evaluation. The

paper is a methodological and analytical contribution; it does not report an empirical deployment or benchmark study.

2. Related Work and Problem Formulation

Reinforcement learning from human feedback has become a dominant approach for aligning language models with preferences [1], [2], [3]. Its success depends on the quality and coverage of the reward model, as well as on the policy optimization procedure. In domain-specific applications, preference labels may be scarce, expensive, or contested, and the reward model may fail to represent rare but severe errors. A policy that maximizes average preference reward can therefore learn to produce answers that are usually acceptable but occasionally catastrophic. The framework developed here retains preference-based utility as one component but adds an explicit risk term and a tail-risk constraint. The purpose is not to replace preference optimization but to prevent average-case utility from dominating rare failure modes.

Retrieval-augmented generation provides external evidence at inference time [4], [5], [6]. REALM-style pretraining and large-scale retrieval show that retrieval can improve knowledge-intensive tasks, while later work has examined how retrieved passages should be integrated into generation [14]. Uncertainty-aware retrieval and generation methods attempt to control risk when retrieval evidence is weak or ambiguous [7]. The present framework extends this direction by making evidence sufficiency a state variable that conditions both risk assessment and action selection. The system does not treat retrieval as a preprocessing step; it can decide to retrieve again, request clarification, or decline to answer when the evidence state is inadequate.

Calibration and uncertainty quantification are central to trustworthy response control [8], [9], [10]. A model's stated confidence may be useful, but it is not a substitute for a calibrated probability of domain-specific error. Hallucination and faithfulness studies show that unsupported generation can survive retrieval and instruction tuning [11], [12], and fact-verification work provides a way to reason about evidential support [13]. The framework therefore combines a learned risk estimator with an evidence verifier. The risk estimator captures task-level consequences, while the verifier captures whether specific claims are supported by retrieved passages. The two signals are related but not identical: a response can be well supported by evidence yet still be risky because the evidence is outdated or because the response exceeds the model's competence.

Formally, let x denote a domain query, E a set of retrieved evidence passages with provenance, a a response action, and Y a candidate response generated under that action. Let A_valid(x,E,a) denote the set of responses that satisfy the domain's acceptability conditions for the query and evidence state. The immediate risk target is the probability that a generated response falls outside this set:

$$R*(x,E,a) = P(Y \notin A_valid(x,E,a) \mid x,E,a)$$

Here P denotes probability under the response model and evidence state, and the indicator Y∉ A_valid means that the response violates at least one acceptability condition. The set A_valid is domain-specific: it may encode factual correctness, jurisdiction, temporal validity, contraindication checks, or citation requirements. This target is not directly observable at inference time, so the framework estimates it with a calibrated risk head rθ and supplements it with a risk-sensitive value critic.

3. Methodological Framework

3.1 Evidence state and sufficiency

The decision state at step t is s_t=(x,E_t,h_t,q_t), where E_t is the current evidence set, h_t is the interaction and retrieval history, and q_t is an evidence-sufficiency score. Each evidence passage carries provenance, timestamp, and verifier outputs indicating support, contradiction, and uncertainty. The sufficiency score is intended to capture whether the evidence covers the factual requirements of the query, independently of how confident the generator is. A proposed formulation is:

$$q(x,E) = \sigma\big(\beta 0 + \beta 1 cov(E) + \beta 2 ver(E) - \beta 3 con(E) - \beta 4 unk(E)\big)$$

In this expression, σ is the logistic function; cov(E) measures coverage of query-required facts by retrieved passages; ver(E) measures positive verification support; con(E) measures contradiction across passages; and unk(E) measures unverifiable or missing evidence. The coefficients β0 through β4 encode the relative importance of these components and may be fixed by domain policy or learned under monotonicity constraints. The score is not a substitute for the risk estimate. Its role is to distinguish a confident answer that is well grounded from a confident answer that is not, and to make evidence insufficiency an explicit reason for retrieval, clarification, abstention, or escalation.

3.2 Reinforcement-learning-based risk assessment

The framework separates immediate risk from cumulative risk. The immediate risk head rθ(x,E,a) estimates the probability that a candidate response generated under action a violates A_valid. It is trained with a calibration objective that encourages the predicted risk to match observed acceptability outcomes when such labels are available. A Brier-style calibration loss is:

$$L_cal(\theta) = E[(r\theta(x,E,a) - 1\{Y \notin A_valid\})\text{^}2]$$

Here 1{·} is the indicator function, and the expectation is over query-evidence-action-response tuples. The loss penalizes both overconfidence and underconfidence, but the consequences of the two errors differ. Underestimation can release unsafe answers, while overestimation can produce excessive abstention. Domain-specific cost asymmetry can be introduced by weighting false-low-risk errors more heavily than false-high-risk errors, but the unweighted form clarifies the calibration target.

Cumulative risk is assessed by a risk-sensitive critic V_r(s) trained alongside the policy. The critic estimates the expected discounted risk cost from a state, allowing the system to account for the fact that an answer may trigger follow-up queries, require correction, or create downstream harm. The actor-critic structure follows standard policy optimization methods [15], [16], but the value target is risk-adjusted rather than purely utility-based. This design connects the risk head to long-horizon response control and avoids treating risk as a single-turn classification problem.

3.3 Risk-constrained policy optimization

The policy π(a|s) selects response actions from the state. The action set includes ANSWER, HEDGE, CITE, RETRIEVE_MORE, ASK_CLARIFY, ABSTAIN, and ESCALATE. These actions have different utilities and costs: retrieval and escalation consume time and human resources, hedging may reduce usefulness, and abstention may be preferable

to an unsupported answer but can also deny needed information. The policy is trained to maximize a risk-adjusted objective with a divergence penalty from a reference policy:

$$J(\pi) = E\pi[\sum t\, \gamma \hat{}\, t(U_t - \lambda r_t - \mu c_t) - \eta KL_t]$$

In this objective, Eπ is the expectation over trajectories induced by π; U_t is domain utility at step t; r_t is the immediate risk estimate; c_t is the action cost; γ is the discount factor; λ weights risk relative to utility; μ weights action cost; and KL_t denotes the Kullback-Leibler divergence between π(·|s_t) and a reference policy. The divergence term limits unstable policy drift and preserves useful behavior learned during instruction tuning or preference optimization.

Average risk control is insufficient when rare errors are severe. The framework therefore adds a conditional value-at-risk constraint on the risk distribution. Let α be the tail probability and ε the permitted risk budget. The constraint is:

$$CVaR\alpha(r\theta) = E[r\theta \mid r\theta \geq VaR\alpha(r\theta)] \leq \varepsilon$$

Here VaRα(rθ) is the value-at-risk threshold at level α, and CVaRα is the expected risk conditional on exceeding that threshold. The constraint forces the policy to account for the upper tail rather than only the mean. Risk-sensitive Markov decision process formulations and CVaR optimization provide the theoretical basis for this construction [18], [19]. In implementation, the constraint can be approximated with a Lagrangian penalty or a distributional critic, but the analytical point is that risk enters the optimization as a constraint on worst-case behavior, not merely as another additive reward term.

3.4 Evidence-aware response control

The trained policy produces a distribution over actions, but deployment requires a deterministic or constrained controller that can enforce domain thresholds. Let r denote the calibrated immediate risk and q the evidence-sufficiency score. The controller maps the pair (r,q) into a response mode:

$$a* = ANSWER\ if\ r \leq \tau L\ and\ q \geq qmin;\ HEDGE\ if\ r \leq \tau H\ and\ q \geq qmin;\ ABSTAIN\ otherwise$$

The thresholds satisfy τL<τH, and qmin is the minimum sufficiency required for a direct answer. In this rule, ANSWER is permitted only when both risk and evidence sufficiency are favorable; HEDGE covers intermediate risk with adequate evidence; and ABSTAIN covers high risk or inadequate evidence. The action set used by the policy is richer than this three-way rule, so an ABSTAIN outcome can be refined into RETRIEVE_MORE, ASK_CLARIFY, or ESCALATE depending on whether the insufficiency is evidential, semantic, or consequential. The threshold structure makes the controller auditable and allows domain risk budgets to be imposed without retraining the language model.

3.5 Risk-adjusted actor update

The actor update uses a risk-adjusted advantage rather than a pure utility advantage. For a transition from state s to s′ under action a, one form is:

$$A\pi(s,a) = U(s,a) - \lambda r\theta(s,a) - \mu c(a) + \gamma V\pi(s') - V\pi(s)$$

Here Aπ(s,a) is the risk-adjusted advantage, Vπ is the value function under policy π, and the remaining symbols follow the objective above. The advantage increases when an action yields high domain utility and low risk cost, and decreases when the action incurs retrieval, latency, or escalation costs. This formulation allows the policy to learn that a hedged or

clarifying response is preferable when the risk estimate is moderate and evidence is incomplete. It also aligns the actor update with the CVaR constraint when the critic is trained to represent the risk tail.

4. Analytical Assessment and Design Trade-offs

4.1 Calibration and threshold sensitivity

The behavior of the controller depends critically on the calibration of r_θ. If the risk estimator systematically underestimates risk, the controller will answer in cases that should be hedged or abstained, and the CVaR constraint will be violated even if the average risk appears acceptable. If it overestimates risk, the system will over-abstain, shifting burden to human escalation and reducing usefulness. The threshold rule makes this trade-off explicit: τ_L controls the boundary between direct answers and hedged answers, τ_H controls the boundary between hedging and abstention, and q_{min} controls how much evidence is required before a direct answer is allowed. Raising τ_H or lowering q_{min} increases coverage but also increases exposure to unsupported responses. Lowering τ_H or raising q_{min} reduces exposure but increases abstention.

The separation of r and q improves sensitivity analysis because the two signals can fail in different ways. A high-confidence generator may produce a fluent answer even when q is low; the controller should not interpret fluency as evidence sufficiency. Conversely, a well-covered evidence set may still yield high risk because the query requires professional judgment or because the evidence is legally or clinically inapplicable. The framework therefore treats risk and sufficiency as complementary dimensions rather than collapsing them into a single confidence score. This distinction is also relevant to conformal calibration, which can provide distribution-free coverage guarantees for risk thresholds under exchangeability assumptions [20]. When domain shift breaks exchangeability, the thresholds require recalibration or domain-specific adjustment.

4.2 Evidence sufficiency and abstention behavior

Abstention is often treated as a failure of the model, but in evidence-intensive domains it can be a correct control action. The framework makes abstention a first-class outcome and links it to evidence sufficiency. When q is below q_{min}, the controller can choose RETRIEVE_MORE if additional retrieval is likely to resolve the insufficiency, ASK_CLARIFY if the query is underspecified, or ESCALATE if the risk is high and human judgment is required. This decision depends on the expected value of additional evidence, not only on the current risk estimate. The policy can learn to avoid unnecessary retrieval when the evidence state is already sufficient, which matters because retrieval and human escalation have real costs.

The sufficiency score also helps diagnose failure modes. Low coverage with high verification support suggests that the available passages support a narrow claim but do not address the full query. High contradiction with moderate coverage suggests that the evidence base is contested. High unverifiability suggests that the relevant facts are absent from the retrieved corpus. Each pattern implies a different response: targeted retrieval, qualification, clarification, or escalation. A single risk score would obscure these distinctions, whereas the state representation preserves them for the policy and the controller.

4.3 Policy stability, reward specification, and action costs

Risk-sensitive reinforcement learning introduces additional sources of instability. The risk head and the critic are learned from finite data, and their errors can be amplified by policy optimization. The KL penalty in the objective limits abrupt changes from the reference policy, but it also constrains adaptation to domain-specific risk patterns. A strong penalty preserves safety behavior learned during alignment but may prevent the policy from learning when abstention is unnecessary. A weak penalty allows rapid adaptation but increases the chance of reward hacking or unstable oscillation between answer and abstain actions.

Action costs are equally important. If retrieval is free and escalation is cheap, the policy may overuse them; if they are expensive, the policy may answer prematurely. The cost term c(a) should therefore reflect operational constraints such as latency, human workload, and the cost of interrupting a user. The risk weight λ and the CVaR budget ε determine how conservatively the policy behaves. Increasing λ or decreasing ε shifts behavior toward hedging, retrieval, and abstention. The appropriate setting is domain-dependent and should be derived from acceptable risk tolerances rather than from a universal default.

4.4 Domain-specific error modes and evaluation logic

Evaluation of the framework should distinguish linguistic quality from control quality. Useful evaluation dimensions include risk calibration, evidence sufficiency calibration, abstention precision and recall, escalation appropriateness, citation faithfulness, and tail-risk behavior. A system that answers most queries correctly but fails on rare high-risk cases may appear strong under average accuracy while being unsuitable for deployment. Evaluation should therefore stratify by risk tier and evidence condition, including queries with missing evidence, contradictory evidence, outdated evidence, and evidence that requires professional judgment.

Domain-specific error modes vary. In clinical decision support, a response may be factually correct in general but contraindicated for a specific patient context. In legal research, a passage may be authoritative in one jurisdiction but irrelevant in another. In financial compliance, a rule may have changed after the retrieved document was published. These errors are not captured by generic hallucination metrics alone. The evidence state should include temporal validity, jurisdiction, and applicability conditions, and the risk head should be trained or specified with these features in mind. The framework's emphasis on provenance and evidence sufficiency is intended to make such conditions visible to the controller.

5. Technical Discussion and Implementation Conditions

The proposed framework is best understood as a control layer around a domain-specific generator and retriever. The generator produces candidate responses; the verifier and sufficiency scorer assess evidence; the risk head and critic assess consequence; and the policy selects a response mode. This decomposition allows components to be updated independently. A new retriever can be introduced without retraining the language model, and a new risk threshold can be imposed without changing the policy. The decomposition also supports audit: each response can be traced to an evidence state, a risk estimate, a sufficiency score, and a selected action.

Implementation requires care in at least three areas. First, the acceptability set A_{valid} must be specified with domain experts. It is not a generic helpfulness or harmlessness label; it encodes the conditions under which a response is admissible for the domain. Second, the risk head must be trained or calibrated on examples that include rare severe errors, or it must be supplemented with rule-based checks for known high-risk conditions. Third, the controller

must be evaluated under distribution shift, because retrieval corpora, regulations, and clinical guidelines change. Alignment training can reduce harmful behavior [21], but it does not replace domain-specific risk assessment and evidence verification.

The framework also has boundary conditions. It assumes that evidence can be retrieved and that provenance is available; it may be less effective in domains where the relevant knowledge is tacit or where evidence is intentionally withheld. It assumes that risk can be estimated from features of the query, evidence, and action; if the risk depends on unobserved user context, the controller should ask for clarification or escalate rather than answer. It assumes that action costs can be meaningfully specified; if they cannot, the policy may learn a distorted trade-off. These conditions do not invalidate the framework, but they determine when it is appropriate and what must be monitored after deployment.

Compared with retrieval-only systems, the framework adds explicit risk assessment and response-mode selection. Compared with preference-optimized models alone, it adds tail-risk constraints and evidence sufficiency. Compared with rule-based abstention, it allows the policy to learn when retrieval, clarification, or escalation is more valuable than refusal. The combination is technically justified because trustworthiness in domain-specific settings depends on both what the model knows and whether it can recognize the limits of what it knows.

6. Conclusion

This manuscript has developed a methodological framework for trustworthy domain-specific language models through reinforcement-learning-based risk assessment and evidence-aware response control. The framework represents queries, evidence, provenance, contradiction signals, and interaction history in a decision state; estimates immediate and cumulative risk; separates evidence sufficiency from answer confidence; and trains a policy to choose among answer, hedge, cite, retrieve, clarify, abstain, and escalate actions under a risk budget. The calibration loss, risk-constrained objective, conditional value-at-risk constraint, threshold rule, and risk-adjusted advantage function form a connected technical design rather than a collection of independent safeguards.

The main analytical insight is that trustworthy response control requires two distinctions that are often collapsed: the distinction between confidence and risk, and the distinction between answer confidence and evidence sufficiency. A confident model can be wrong, and a well-grounded model can still face high-consequence uncertainty. The proposed controller uses both dimensions to decide whether to answer, qualify, retrieve, or escalate. The discussion of calibration sensitivity, action costs, reward specification, and domain-specific error modes shows where the framework is likely to succeed and where it requires domain validation. The approach is intended for evidence-intensive deployments in which abstention and escalation are legitimate outcomes and in which tail risk matters as much as average performance.

References

1. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

2. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30, 4299–4307.

3. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008–3021.

4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

5. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). REALM: Retrieval-augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning (pp. 3929–3938). PMLR.

6. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., … Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In Proceedings of the 39th International Conference on Machine Learning (pp. 2206–2240). PMLR.

7. Zhang, R. (2025, December). Uncertainty-Aware RAG: Trustworthy Generation and Risk Control. In 2025 5th International Symposium on Artificial Intelligence and Big Data (AIBDF) (pp. 526-531). IEEE.

8. 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:2207.05221.

9. Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. arXiv:2205.14334.

10. 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 (pp. 1321–1330). PMLR.

11. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

12. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1906–1919). Association for Computational Linguistics.

13. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In Proceedings of NAACL-HLT 2018 (pp. 809–819). Association for Computational Linguistics.

14. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997.

15. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv:1707.06347.

16. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the 35th International Conference on Machine Learning (pp. 1861–1870). PMLR.

17. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv:1606.06565.

18. Chow, Y., & Ghavamzadeh, M. (2014). Algorithms for CVaR optimization in MDPs. Advances in Neural Information Processing Systems, 27.

19. Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. Journal of Risk, 2(3), 21–42.

20. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511.

21. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.