

Multimodal Uncertainty-Aware RAG for Autonomous Driving Perception and Safety-Critical Decision-Making

Kari Rocha

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.

kari.rocha643@ku.edu

Waida Varma

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.

waidawork@uc.edu

Abstract

Autonomous driving systems increasingly depend on multimodal perception, language-capable reasoning, and retrieval-augmented generation to interpret complex environments and support safety-critical decisions. Yet the integration of retrieval mechanisms into vehicular perception and planning introduces new forms of epistemic, aleatoric, and infrastructure-mediated uncertainty that are not fully addressed by conventional sensor fusion or end-to-end learning pipelines. This paper examines multimodal uncertainty-aware retrieval-augmented generation as a system-level architecture for autonomous driving perception and safety-critical decision-making. It argues that uncertainty must be treated not merely as a model output but as a cross-layer property spanning sensors, embeddings, retrieval indexes, generative reasoning, runtime monitors, governance structures, and fleet operations. The analysis considers structural trade-offs among latency, interpretability, robustness, redundancy, energy consumption, and regulatory accountability. It also explores how uncertainty-aware retrieval can support scenario interpretation, rare-event handling, and human-readable explanations while creating risks related to stale knowledge, retrieval bias, hallucinated grounding, and opaque escalation policies. The paper further discusses deployment constraints at the edge, lifecycle sustainability, fairness in geographic and demographic representation, and policy implications for certification and liability. It concludes that multimodal uncertainty-aware RAG can improve autonomous driving safety only when retrieval, generation, and decision-making are governed by calibrated uncertainty thresholds, fail-operational fallback logic, auditable data provenance, and continuous operational learning. The contribution is a conceptual and infrastructural roadmap for researchers, engineers, and policymakers seeking trustworthy multimodal intelligence in safety-critical mobility systems.

Keywords

multimodal perception; retrieval-augmented generation; uncertainty quantification; autonomous driving; safety-critical decision-making; runtime assurance; AI governance; edge computing.

1. Introduction

Autonomous driving has advanced through improvements in sensor hardware, deep perception models, high-definition mapping, and increasingly capable planning architectures. Transformer-based sequence modeling has provided a general computational substrate for integrating heterogeneous signals and long-range dependencies [1], while vision transformers and multimodal contrastive learning have extended representation learning across images, text, and other modalities [2], [3], [4]. These developments have made it possible to imagine vehicles that not only detect objects but also retrieve contextual knowledge, reason over traffic rules, and generate explanations for their decisions. However, safety-critical mobility differs from conventional machine learning benchmarks because errors are embodied, consequential, and often irreversible. A perception model that is uncertain about a pedestrian, a retrieval system that returns an outdated regulation, or a generative planner that produces a plausible but unsupported maneuver can all contribute to harm. The central problem is therefore not simply better prediction but disciplined uncertainty management across the entire socio-technical stack.

Retrieval-augmented generation has emerged as a promising method for grounding language and multimodal models in external knowledge [11]. Early retrieval-augmented language models demonstrated that access to non-parametric memory can improve knowledge-intensive tasks [12], and subsequent work has refined passage retrieval, generative conditioning, and large-scale retrieval from massive corpora [13], [14]. In autonomous driving, similar mechanisms could retrieve map priors, traffic codes, similar driving scenarios, maintenance records, or human-labeled edge cases. A vehicle encountering an unusual construction zone might retrieve relevant municipal regulations, prior annotated scenes, or fleet-level incident reports. A planner facing ambiguous lane markings might retrieve map provenance, weather advisories, or near-miss cases from comparable locations. The appeal is clear: retrieval can make the system more adaptive, explainable, and responsive to long-tail conditions without retraining the entire model.

Yet retrieval also introduces uncertainty that is qualitatively different from sensor noise. The relevance of a retrieved document may be uncertain, the embedding may misrepresent semantic similarity, the index may be stale, and the generator may overstate the authority of retrieved content. Recent work on uncertainty-aware retrieval-augmented generation emphasizes trustworthy generation and risk control, suggesting that retrieval systems need explicit mechanisms for estimating when retrieved evidence is sufficient, conflicting, or unsafe [15]. This paper develops that insight for autonomous driving by treating multimodal uncertainty-aware RAG as a layered architecture rather than a single model. It examines how uncertainty can be represented across sensors, perception, retrieval, generation, planning, and governance. It then analyzes structural trade-offs in deployment, including latency, redundancy, energy, fairness, and auditability. The argument is that autonomous driving systems should not merely use retrieval to improve average-case performance; they must use uncertainty-aware retrieval to decide when to trust, defer, escalate, or refuse a generated recommendation.

2. Foundations of Multimodal Perception and Retrieval-Augmented Generation

Multimodal perception in autonomous driving combines cameras, lidar, radar, ultrasonic sensors, inertial measurement units, and sometimes vehicle-to-everything communication. Each modality has distinct failure modes, spatial resolutions, temporal characteristics, and vulnerability to environmental conditions. Cameras provide rich semantic detail but degrade in glare, fog, or darkness. Lidar offers precise geometry but can be affected by rain, dust, and

reflective surfaces. Radar is robust in poor weather but has lower angular resolution. A multimodal system therefore does not simply concatenate data; it must learn when modalities agree, when they conflict, and when one should dominate. Transformer architectures have made such fusion more flexible by allowing attention across modalities and time [1], while vision transformers and contrastive pretraining provide scalable representations that can align visual, textual, and sensor-derived embeddings [2], [3], [4]. These representations are essential for retrieval because they allow a vehicle to query knowledge using multimodal cues rather than text alone.

Retrieval-augmented generation builds on the idea that generative models should not rely solely on parametric memory. In natural language processing, retrieval-augmented generation combines a retriever with a generator to produce answers grounded in external documents [11]. REALM demonstrated that retrieval can be integrated into pretraining, allowing models to access a large corpus during learning [12]. Later work improved open-domain question answering by leveraging passage retrieval with generative models [13], and large-scale retrieval from trillions of tokens showed that non-parametric memory can substantially expand model knowledge [14]. For autonomous driving, the retrieved knowledge base could include traffic laws, map metadata, operational design domain constraints, maintenance histories, simulation scenarios, and annotated incident reports. The generator could then produce a decision rationale, a maneuver proposal, or a query to a human supervisor. This architecture is attractive because it separates knowledge from model weights, enabling updates without full retraining.

However, multimodal RAG in vehicles is not a straightforward transfer from text-based question answering. Driving decisions are continuous, time-sensitive, and physically constrained. A retrieved regulation may be relevant in general but inapplicable to a specific intersection due to local signage, temporary construction, or weather. A retrieved scenario may be semantically similar but dynamically different because of speed, occlusion, or road friction. A generator may produce a coherent explanation that does not correspond to the actual planning constraints. The system must therefore estimate uncertainty at multiple points: sensor reliability, object detection confidence, semantic embedding ambiguity, retrieval relevance, evidence consistency, generator calibration, and planner risk. The required integration is not merely multimodal but multi-source and multi-timescale, linking real-time perception with slower-moving institutional knowledge. This is where uncertainty-aware RAG becomes a systems problem rather than a component-level optimization.

3. Uncertainty Representation Across Sensors, Models, and Knowledge Infrastructures

Uncertainty in autonomous driving is often divided into aleatoric and epistemic forms. Aleatoric uncertainty arises from inherent noise in observations, such as sensor degradation or ambiguous object boundaries. Epistemic uncertainty arises from limited knowledge, such as encountering a novel scenario or an out-of-distribution object. Bayesian deep learning has provided frameworks for distinguishing these uncertainties in computer vision, especially for safety-critical perception [5]. Practical approximations such as dropout-based Bayesian inference and deep ensembles have made uncertainty estimation more scalable [6], [7]. These methods are valuable because they allow a perception system to signal not only what it sees but how confident it is. In a multimodal RAG architecture, such signals should propagate into retrieval and generation rather than being discarded at the perception boundary.

Retrieval introduces a third form: knowledge uncertainty. This includes uncertainty about whether the retrieved document exists, whether it is current, whether it is authoritative,

whether it applies to the current jurisdiction, and whether the embedding space correctly represents the query. A vehicle may retrieve a traffic rule from a neighboring state, a map segment from a previous construction period, or an incident report from a different vehicle platform. The generator may then blend this evidence into a confident recommendation. Uncertainty-aware retrieval must therefore maintain provenance, timestamps, geographic scope, and confidence metadata for each knowledge item. It must also detect conflicts among retrieved sources. If a retrieved law conflicts with a live traffic signal, the system should privilege the live signal while recording the discrepancy for later review. If two retrieved scenarios suggest different maneuvers, the planner should treat the disagreement as a risk signal rather than averaging them into false consensus.

Generative uncertainty adds another layer. Large language and multimodal models can produce fluent but unsupported claims, a problem often described as hallucination. In safety-critical driving, hallucination is not merely an informational error; it can become a maneuver. Uncertainty-aware generation should therefore expose whether a recommendation is grounded in retrieved evidence, inferred from model priors, or generated from conflicting sources. A recent framework for uncertainty-aware retrieval-augmented generation emphasizes trustworthy generation and risk control, indicating that confidence calibration and refusal behavior are central to deployment [15]. For autonomous vehicles, refusal may mean falling back to a conservative policy, requesting remote assistance, or limiting operational speed. The key architectural insight is that uncertainty should be represented in a form that planners, runtime monitors, and human supervisors can act upon. A scalar confidence score is insufficient; the system needs structured uncertainty that identifies the source, scope, and consequence of doubt.

4. System Architecture and Structural Trade-Offs

A multimodal uncertainty-aware RAG architecture for autonomous driving can be understood as a layered pipeline with feedback loops. At the lowest layer, sensors produce observations with quality metrics. The perception layer fuses these observations into object tracks, free-space estimates, semantic maps, and scene graphs. The embedding layer converts multimodal scene representations into queries that can be matched against local and remote knowledge indexes. The retrieval layer returns evidence with provenance, relevance scores, and uncertainty estimates. The generation layer synthesizes explanations, rule interpretations, and decision options. The planning layer evaluates those options against physical dynamics, safety envelopes, and passenger comfort. The runtime assurance layer monitors the entire chain for inconsistencies, latency violations, and degraded modes. Finally, the governance layer manages data provenance, updates, audits, and liability. This layered view reveals that uncertainty-aware RAG is not a single module but a coordination mechanism across the vehicle, fleet, cloud, and regulatory environment.

The first structural trade-off is latency versus depth. Retrieval from a large cloud index can provide rich contextual knowledge, but communication delays, bandwidth limits, and network outages make cloud-only retrieval unsafe. Edge retrieval from a local cache can reduce latency, but the cache may be stale or incomplete. A hybrid design can retrieve safety-critical local knowledge immediately while asynchronously updating broader knowledge when connectivity permits. The second trade-off is interpretability versus performance. Larger generative models may produce better explanations and more flexible reasoning, but they are harder to audit and calibrate. Smaller models may be more predictable but less capable of handling novel scenarios. The third trade-off is redundancy versus energy. Multiple sensors,

multiple retrievers, and multiple planners can improve robustness, but they increase weight, power consumption, and computational heat. Electric and autonomous vehicles already face energy constraints, so uncertainty-aware architectures must justify their resource costs in safety terms.

A fourth trade-off concerns centralized versus distributed knowledge. Centralized indexes allow fleet-wide updates and consistent governance, but they create single points of failure and raise privacy concerns. Distributed indexes at the edge or within regional fleets can improve resilience and jurisdictional relevance, but they complicate version control and certification. A fifth trade-off involves retrieval granularity. Retrieving entire documents may preserve context but overwhelm the generator. Retrieving short passages may improve precision but lose legal nuance. Retrieving structured rules may support formal reasoning but struggle with ambiguous human language. These trade-offs cannot be resolved universally; they depend on operational design domain, regulatory regime, vehicle platform, and risk tolerance. The system architecture should therefore make these trade-offs explicit and configurable rather than hiding them inside a monolithic model.

5. Safety-Critical Decision-Making and Runtime Assurance

Safety-critical decision-making requires more than high accuracy. It requires bounded behavior under uncertainty. Autonomous vehicle safety frameworks often emphasize safe fallback states, minimal risk maneuvers, and verifiable constraints [19]. Koopman and Wagner argue that autonomous vehicle safety is an interdisciplinary challenge involving engineering, human factors, ethics, and regulation [20]. In a RAG-enabled vehicle, the planner should not accept a generated recommendation solely because it is fluent. It should require that the recommendation is supported by retrieved evidence whose uncertainty is below a context-specific threshold, that it satisfies physical and legal constraints, and that it remains safe under plausible perception errors. If these conditions are not met, the system should enter a degraded mode. Runtime assurance can monitor whether retrieved knowledge is current, whether the generator is within its calibration domain, and whether the planner is relying on assumptions that conflict with live sensor data.

One important design pattern is the separation of generative reasoning from safety-critical control. The generative component can propose interpretations, explanations, and high-level maneuver options, but a separate verifier should check them against hard constraints. This verifier may use formal methods, rule-based logic, reachability analysis, or conservative control barriers. The verifier should also consider uncertainty from retrieval. For example, if the vehicle retrieves a rule permitting a right turn on red, the verifier should confirm that the local jurisdiction, signage, and signal state support that rule. If the evidence is uncertain, the verifier should default to a more conservative behavior. This separation reduces the risk that a hallucinated or outdated retrieval directly actuates the vehicle. It also creates a clearer audit trail: the generative system explains why it proposed a maneuver, and the verifier records why it accepted, modified, or rejected that proposal.

Another pattern is uncertainty-triggered escalation. When uncertainty exceeds a threshold, the vehicle may reduce speed, increase following distance, activate hazard lights, request remote assistance, or pull over. Escalation should be graduated rather than binary. Low uncertainty might allow normal operation with logging. Moderate uncertainty might trigger conservative driving and additional sensing. High uncertainty might trigger minimal risk condition. This graduated response requires a mapping from uncertainty sources to operational consequences. Sensor uncertainty might be mitigated by slowing down. Retrieval uncertainty might be

mitigated by querying additional sources. Generative uncertainty might be mitigated by asking for human confirmation. The mapping should be designed, validated, and certified rather than learned implicitly. In this sense, uncertainty-aware RAG is not only a perception or language technology; it is a safety architecture that determines when the vehicle should trust its own reasoning.

6. Governance, Fairness, and Policy Implications

Governance is essential because multimodal RAG systems depend on knowledge infrastructures that extend beyond the vehicle. Traffic laws, map data, incident reports, and maintenance records are produced by institutions with different standards, update cycles, and legal authorities. If a vehicle retrieves outdated or incorrect knowledge, responsibility may be distributed across manufacturers, map providers, municipalities, and data curators. Clear provenance and versioning are therefore necessary for liability and accountability. Policy frameworks may need to specify how retrieval sources are certified, how uncertainty is reported, and how updates are validated before deployment. Foundation model research has already highlighted the risks of large-scale pretrained systems, including opacity, bias, and concentration of power [10]. These risks are amplified when foundation models are combined with retrieval and embedded in public roads.

Fairness is another concern. Retrieval indexes may underrepresent rural roads, informal settlements, minority languages, or regions with limited digital mapping. A vehicle operating in a well-mapped city may receive rich contextual knowledge, while a vehicle in a poorly mapped area may face higher uncertainty and more conservative behavior. This disparity can translate into unequal mobility, economic exclusion, or safety gaps. Fairness also applies to scenario retrieval. If incident reports are dominated by certain vehicle types, weather conditions, or demographic contexts, the system may fail to generalize to underrepresented conditions. Governance should therefore require representative data collection, transparent coverage metrics, and mechanisms for affected communities to report errors. Uncertainty-aware RAG can help by exposing low-confidence regions, but exposure alone is not sufficient; institutions must act on those signals.

Policy implications also include privacy and cybersecurity. Retrieval queries may reveal vehicle location, passenger behavior, or infrastructure vulnerabilities. A compromised retrieval index could inject malicious rules or misleading scenarios. A compromised generator could produce unsafe explanations that appear authoritative. Regulatory approaches may need to treat retrieval infrastructure as safety-critical software, subject to security audits, redundancy requirements, and incident reporting. Cross-border operations raise additional challenges because legal knowledge is jurisdictional. A vehicle crossing a state or national boundary must update its retrieval sources, reconcile conflicting rules, and verify that its planner respects local constraints. This suggests that multimodal uncertainty-aware RAG should be governed as part of a broader mobility data ecosystem, not as an isolated AI feature. The technical architecture and the policy architecture must co-evolve.

7. Deployment, Sustainability, and Lifecycle Operations

Deployment of multimodal uncertainty-aware RAG at scale involves significant infrastructure. Vehicles need onboard compute for real-time perception and generation, memory for local indexes, and communication links for updates. Cloud or regional data centers may host large indexes, model training, and fleet analytics. Edge servers at roadside units or depots may provide low-latency retrieval for specific zones. This hybrid topology must be managed

across the lifecycle. Models and indexes require versioning, rollback, monitoring, and retirement. A retrieval item that was valid during a construction period may become misleading after construction ends. A generative model that was calibrated on one sensor suite may drift when a new lidar is installed. Lifecycle operations must therefore treat uncertainty calibration as a continuous process rather than a one-time validation.

Sustainability is closely tied to deployment choices. Larger models and richer retrieval indexes consume more energy during training and inference. Continuous cloud communication adds network energy and may strain bandwidth in dense urban areas. Redundant sensors and compute increase vehicle mass, which affects range and tire wear. These costs must be weighed against safety benefits. A sustainable design might use tiered retrieval, where simple queries are handled locally and only ambiguous cases invoke cloud resources. It might use model distillation, quantization, and sparse activation to reduce inference energy. It might also use federated learning to update models without centralizing sensitive data. However, efficiency optimizations can reduce uncertainty calibration if not carefully evaluated. The goal is not minimal computation but justified computation: resources should be allocated where they most reduce risk.

Operational sustainability also includes maintenance and human oversight. Remote assistance centers, safety operators, and data annotation teams are part of the system. Their workload depends on how often the vehicle escalates. If uncertainty thresholds are too low, the vehicle may request help constantly, creating fatigue and scaling problems. If thresholds are too high, the vehicle may take unnecessary risks. Calibration therefore has an organizational dimension. Fleet managers must monitor escalation rates, false alarms, missed detections, and regional 差异. They must also ensure that human supervisors have enough context to make good decisions. A well-designed RAG system can provide concise evidence summaries, uncertainty indicators, and alternative maneuver options. A poorly designed system can overwhelm supervisors with verbose explanations or hide the most important risk factors. Deployment success depends on aligning technical uncertainty with human decision capacity.

8. Evaluation, Robustness, and Cross-Domain Lessons

Evaluating multimodal uncertainty-aware RAG for autonomous driving requires more than standard perception metrics. Datasets such as KITTI, nuScenes, and the Waymo Open Dataset have driven progress in multimodal perception and tracking [16], [17], [18]. They provide benchmarks for detection, segmentation, and motion prediction, but they do not fully capture retrieval uncertainty, knowledge conflicts, or generative grounding. Evaluation must therefore include scenario-based testing, adversarial retrieval, stale-index simulations, and human-in-the-loop studies. Robustness testing should examine sensor degradation, network outage, map error, legal conflict, and malicious injection. It should also measure calibration: when the system reports high confidence, how often is it correct? When it reports high uncertainty, does it actually escalate safely? These questions require metrics that connect uncertainty to operational outcomes.

Cross-domain lessons can inform autonomous driving. In healthcare, retrieval-augmented systems must ground clinical recommendations in guidelines while recognizing patient-specific context. In aviation, checklists and redundant systems manage uncertainty through procedure and governance. In finance, risk models use stress testing and circuit breakers to prevent cascading failures. Each domain suggests that uncertainty-aware RAG should not be evaluated only as a language task. It should be evaluated as a decision-support system with

failure modes, escalation policies, and accountability. The autonomous driving context is especially demanding because the system operates in public space with diverse actors. A retrieval error that would be a minor inconvenience in a chatbot can become a traffic hazard. Therefore, evaluation should include rare events, long-tail scenarios, and cross-jurisdictional rule conflicts.

A further lesson is that robustness and fairness are linked. A system that is robust only in well-mapped, high-income areas is not robust in a societal sense. A system that is fair in average performance but uncertain in underrepresented contexts may still produce disparate safety outcomes. Uncertainty-aware RAG can contribute to fairness by making coverage gaps visible and by enabling conservative behavior where knowledge is weak. But it can also reinforce inequality if conservative behavior becomes a permanent limitation for certain regions. The design challenge is to combine short-term caution with long-term knowledge expansion. Fleet operations should use uncertainty signals to prioritize data collection, mapping, and validation in underserved areas. In this way, uncertainty is not only a safety mechanism but also a learning signal for equitable infrastructure development.

9. Conclusion

Multimodal uncertainty-aware retrieval-augmented generation offers a promising direction for autonomous driving perception and safety-critical decision-making, but its value depends on systemic discipline. Retrieval can enrich perception with contextual knowledge, support rare-event reasoning, and generate human-readable explanations. Uncertainty estimation can signal when sensors, embeddings, retrieved evidence, or generative outputs should not be trusted. Yet these benefits do not emerge automatically from combining multimodal models with a knowledge index. They require an architecture that propagates uncertainty across layers, separates generative proposals from safety verification, and provides runtime assurance that can escalate or fall back when confidence is insufficient. They also require governance that manages provenance, versioning, fairness, privacy, cybersecurity, and liability across institutions.

The paper has argued that uncertainty in this setting is not a single quantity but a structured, multi-source property. Sensor uncertainty, model uncertainty, retrieval uncertainty, generative uncertainty, and institutional uncertainty interact in ways that can compound or mask one another. A vehicle may be confident in perception but retrieving irrelevant law, or uncertain in perception but supported by robust local rules. The system must therefore reason about the source and consequence of doubt, not merely its magnitude. Structural trade-offs among latency, interpretability, redundancy, energy, and centralization must be made explicit and calibrated to operational design domains. Deployment and sustainability considerations must include edge computing, lifecycle updates, human oversight, and energy justice. Evaluation must extend beyond benchmark accuracy to calibration, escalation behavior, adversarial robustness, and equitable coverage.

Ultimately, multimodal uncertainty-aware RAG should be understood as a socio-technical safety infrastructure. Its success will depend on technical advances in calibration, retrieval, and multimodal representation, but also on policy frameworks that certify knowledge sources, audit updates, and assign responsibility. Autonomous vehicles will operate in a world of incomplete, changing, and contested information. A system that acknowledges this condition and responds with calibrated caution, transparent reasoning, and accountable governance is more likely to earn public trust than one that projects unwarranted certainty. The research agenda should therefore prioritize uncertainty-aware architectures that can refuse, defer, and

learn, while remaining auditable and fair. If developed with these commitments, multimodal RAG can contribute to autonomous driving that is not only intelligent but also trustworthy, sustainable, and socially legitimate.

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. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.
3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748-8763.
4. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, 119, 1597-1607.
5. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 5574-5584.
6. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 48, 1050-1059.
7. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 6402-6413.
8. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
9. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
10. 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*.
11. 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.

12. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning*, 119, 3929-3938.
13. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 874-880.
14. 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., Driessche, G., Hennigan, T., Huang, S., Maggiore, L., Mankowitz, D., ... & Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. *Proceedings of the 39th International Conference on Machine Learning*, 162, 2206-2240.
15. 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.
16. Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3354-3361.
17. Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621-11631.
18. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., ... & Anguelov, D. (2020). Scalability in perception for autonomous driving: Waymo open dataset. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2446-2454.
19. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*.
20. Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, 9(1), 90-96.