

AI-to-AI Communication: Toward a New Paradigm of Intelligence Search

Hamid Shahrokh Shahraki

KFJ Company

h_shahrokh_sh@yahoo.com

Abstract

The rapid development of artificial intelligence has transformed the way people search for and interact with information. Traditional search engines primarily retrieve documents, web pages, and other information sources, while modern large language models can interpret questions, reason over information, and generate answers. This paper explores a possible next step: a search paradigm in which artificial intelligence systems communicate with other AI systems to search for, compare, evaluate, and synthesize answers.

The paper introduces the concept of AI-to-AI Communication for Intelligence Search, in which a user's question is distributed among multiple specialized AI systems rather than being answered by a single model. The participating systems may independently reason about the problem, retrieve external information, critique other responses, and contribute evidence to a final synthesis layer. The proposed framework is conceptually related to retrieval-augmented generation, AI agents, and multi-agent systems, but shifts the focus from retrieving information *for an AI* toward retrieving and evaluating *intelligence from multiple AIs*.

Therefore, the proposed architecture moves beyond answer-level agreement toward claim-level verification. Individual AI responses are decomposed into claims, evidence is retrieved independently, sources are evaluated for independence, and correlated errors are explicitly considered before a final answer is produced.

Based on this perspective, AI-to-AI Intelligence Search can be viewed as a potential new layer between artificial intelligence systems: a layer for discovering, coordinating, evaluating, and verifying distributed machine intelligence.

Keywords

AI-to-AI communication; Intelligence Search; AI agents; agent discovery; multi-agent systems; model routing; multi-LLM collaboration; information retrieval; artificial intelligence

1. Introduction

For decades, search engines have been built around a simple idea: a user has a question, a search engine searches a collection of information, and the system returns documents or web pages that may contain the answer.

This model changed significantly with the development of large language models (LLMs). Instead of simply locating information, an LLM can interpret a question, combine information from its learned parameters, reason about a problem, and produce a natural-language response. However, the knowledge available to a single model remains limited, and models can generate incorrect or unsupported information. Retrieval-augmented generation (RAG) was introduced as one important approach for combining model-based generation with external information retrieval [1]. ([NeurIPS Papers](#))

At the same time, research has increasingly explored systems in which multiple AI agents cooperate. Frameworks such as AutoGen demonstrate that multiple language-model-based agents can communicate and collaborate to solve complex tasks [2]. ReAct similarly demonstrated the value of combining reasoning with interaction with external environments and information sources [3]. ([arXiv](#))

Figure 1 summarizes a broad evolution. Early keyword search primarily matched terms and returned documents. Semantic search improved retrieval by representing meaning and context. Generative AI shifted the interaction toward direct answer generation. RAG connected generation with external information retrieval [1]. Agentic search added planning, tool use, and iterative interaction, while multi-agent systems demonstrated that several specialized agents can work together [2].

These developments suggest an interesting question:

What if an AI system did not search only for information, but searched among other AI systems for the best available intelligence?

This question motivates the concept proposed in this paper: AI-to-AI Communication for Intelligence Search.

The basic idea is not that one AI must become universally intelligent. Instead, a search system could communicate with several AI models or specialized AI agents, ask them to independently analyze a problem, compare their responses, identify disagreements, retrieve additional evidence when necessary, and synthesize a final answer.

Such a system would therefore create a new layer between traditional information retrieval and artificial general intelligence (AGI).

The literature contains strong pieces of the proposed vision, but these pieces are usually studied under different names and objectives. RAG focuses on information retrieval for generation. Agentic systems focus on planning and action. Multi-agent systems focus on coordination. Model routing focuses on selecting an appropriate model. Agent discovery focuses on finding agents. Multi-LLM collaborative search focuses on using multiple LLMs for complex search and reasoning; recent work such as MOSA demonstrates the potential of iterative collaboration among multiple LLMs [1- 8].

The gap addressed here is therefore conceptual rather than a claim of technological absence. We propose a unified perspective in which the system searches for task-relevant intelligence, communicates with selected AI systems, evaluates their contributions, and synthesizes the resulting intelligence.

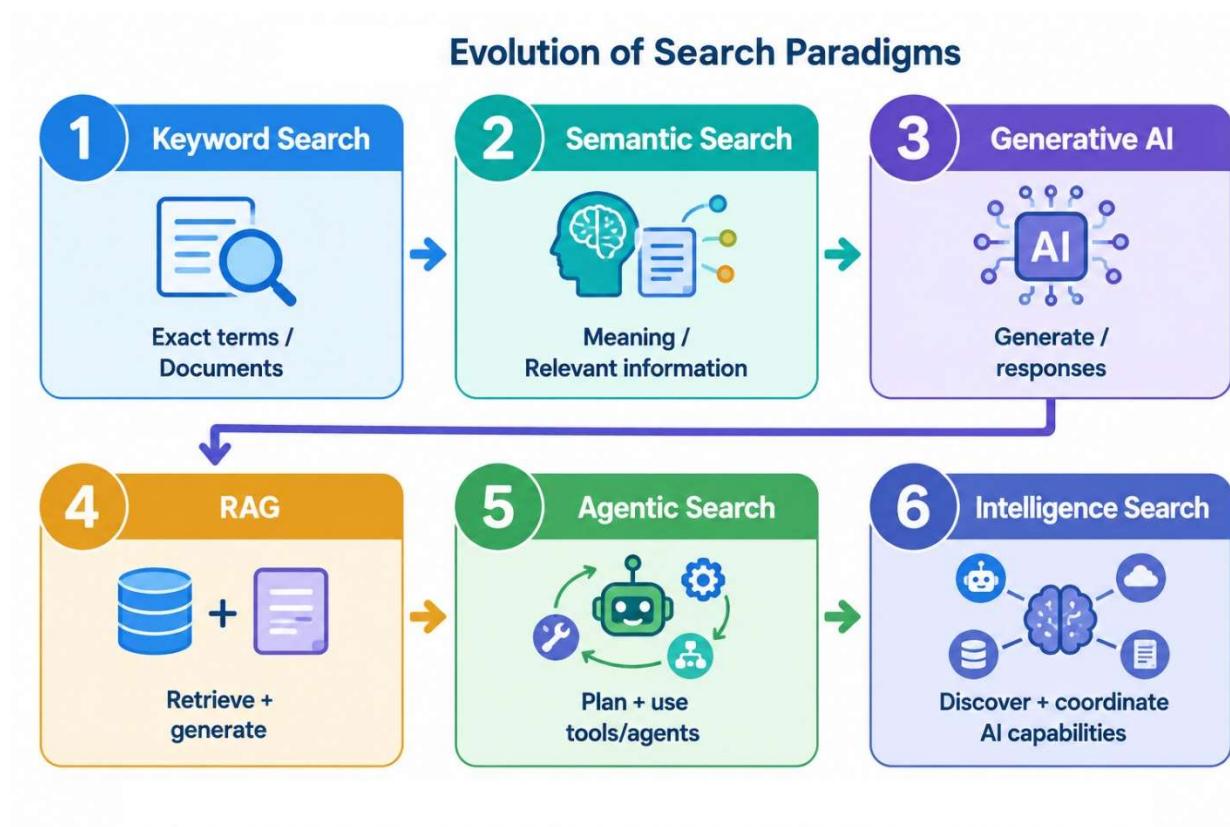


Figure 1. Evolution of Search Paradigms.

2. From Information Search to Intelligence Search

Figure 2 contrasts conventional information search with the intelligence Search perspective. The difference begins with the input. A traditional search system receives a question or query. An Intelligence Search system receives a task, goal, or problem together with constraints such as required expertise, reliability, latency, cost, or available tools.

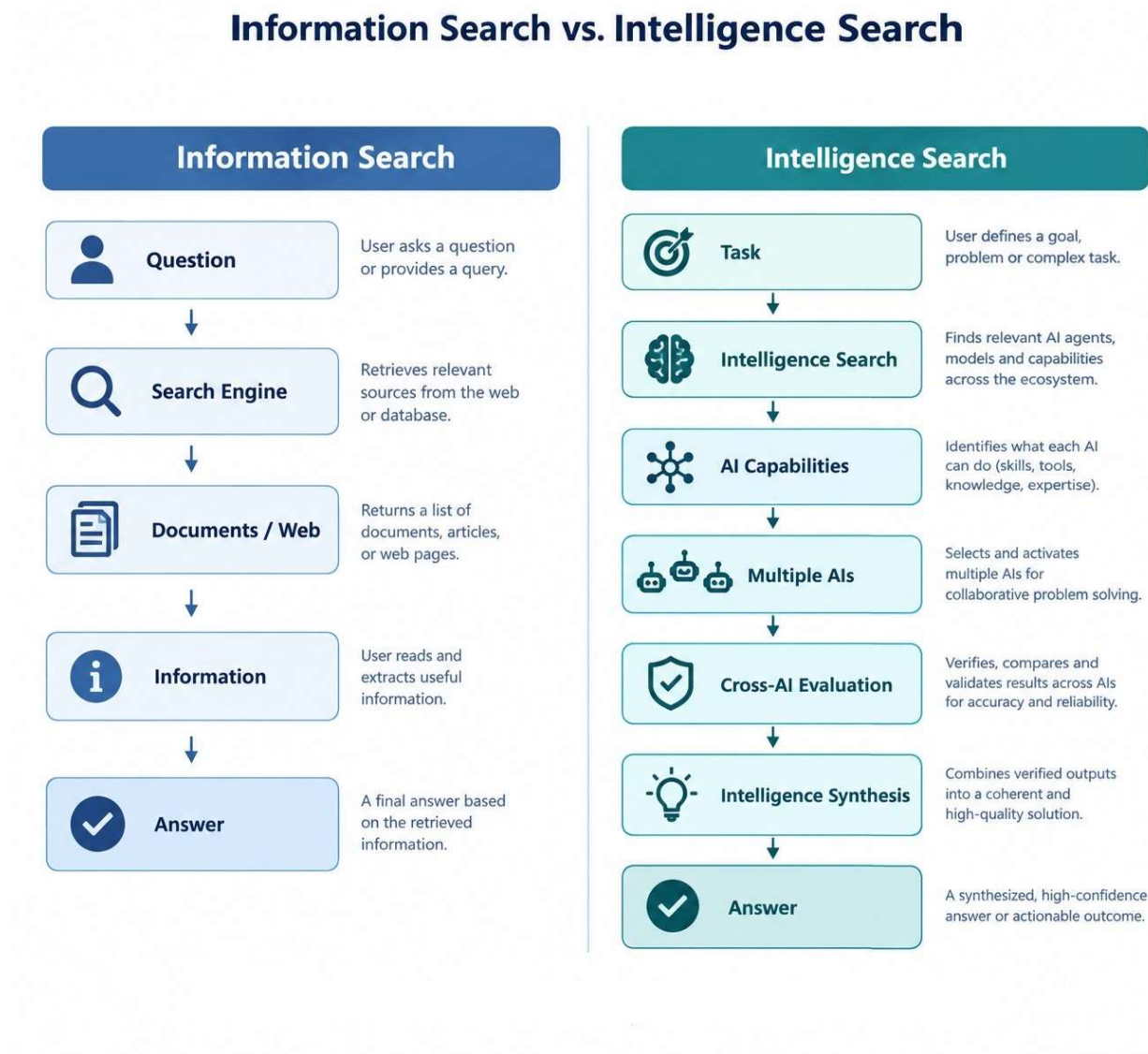


Figure 2. Information Search versus Intelligence Search.

Traditional search engines primarily operate on information objects such as documents, web pages, images, videos, and databases. In an AI ecosystem, however, useful computational resources may also include language models, specialist agents, reasoning modules, tool-enabled agents, databases, simulators, and verification systems. These resources can be selected according to their capabilities.

This leads to a distinction between three levels of search. Information Search retrieves information. Agent Search retrieves an agent that can execute a task. Intelligence Search retrieves and coordinates the capabilities needed to solve a task, potentially using several agents or models and evaluating their outputs before producing a final result.

As previously mentioned, a wide range of research efforts in the field of intelligence search are currently underway and continue to evolve, each approach addresses particular aspects of the problem and has its own limitations. Retrieval-Augmented Generation (RAG), for example, improves access to external knowledge but does not, by itself, guarantee that retrieved information or generated answers are correct [1]. Similarly, multi-agent frameworks such as AutoGen facilitate interaction among AI agents, yet coordination and communication alone do not ensure factual agreement or reliable outcomes [2]. ReAct combines reasoning with external actions, but its performance remains dependent on the quality of the underlying model, the available tools, and the information obtained during execution [3]. Multi-model search systems, including LLM-MS, further explore the use of multiple language models for information discovery, while recent collaborative-search research investigates how different reasoning paths can improve complex problem-solving [4, 8]. However, using more AI models does not necessarily produce a more accurate final answer, particularly when models share similar biases, rely on overlapping sources, or reinforce one another's errors. Moreover, agent discovery and retrieval introduce an additional challenge: selecting an agent based on its description or apparent relevance does not necessarily establish its actual effectiveness for a given task [5].

These limitations highlight the need to move beyond simply retrieving, combining, or coordinating AI-generated responses toward a more systematic approach that evaluates the evidence, independence, reliability, and task-specific performance of participating AI systems. This need motivates further investigation of AI-to-AI Communication as a potential paradigm for intelligence search, in which the quality of the final answer depends not merely on the number of participating models, but on how their contributions are assessed, compared, and validated.

3. Proposed Architecture

Figure 3 presents the proposed architecture as a complete pipeline. A user task is transformed into a required intelligence profile. Intelligence Search discovers candidate AI systems and selects a suitable set. These systems communicate and collaborate. Their outputs are evaluated, agreements and conflicts are identified, and validated insights are fused into a final answer.

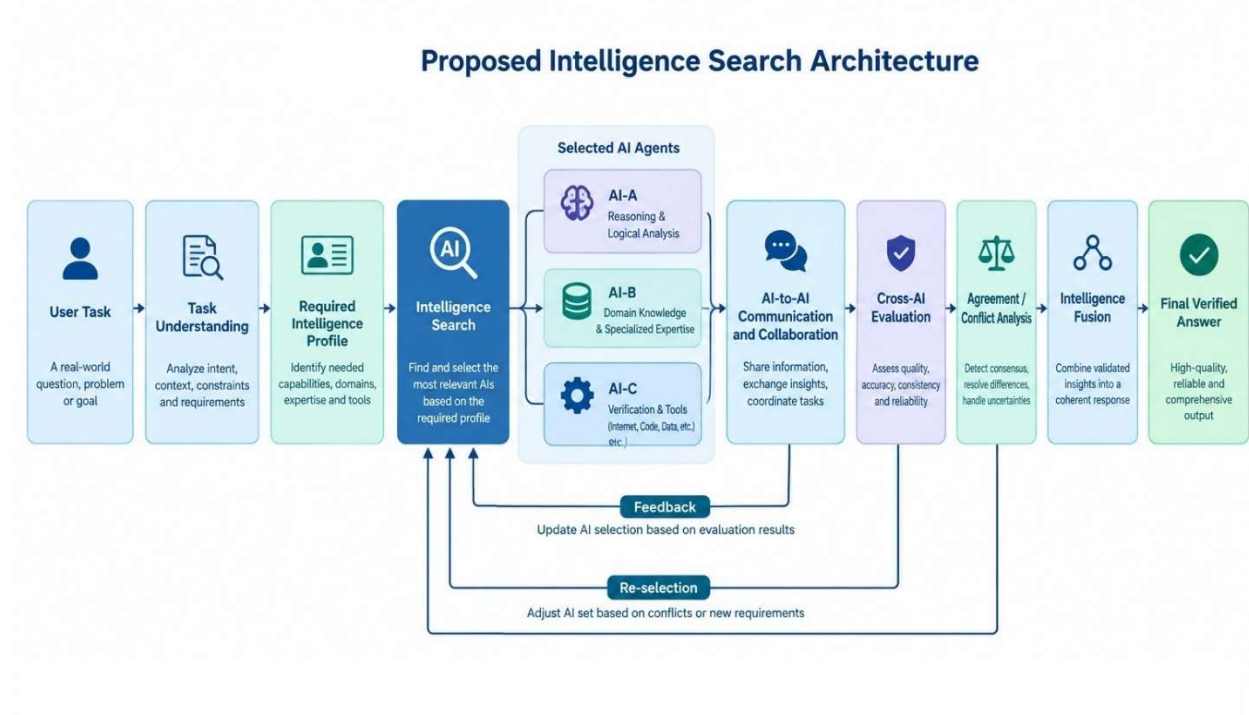


Figure 3. Proposed Intelligence Search Architecture.

3.1 Task Understanding

The system first analyzes the task and identifies its intent, complexity, constraints, and expected output. A complex task may require several different forms of intelligence. For example, a scientific question may require literature retrieval, mathematical reasoning, domain expertise, data analysis, and independent verification.

3.2 Required Intelligence Profile

The task is converted into an intelligence profile. This profile can include required capabilities, preferred domains, tool requirements, reliability targets, cost limits, response-time constraints,

and the desired degree of independent verification. The profile becomes the query used by the Intelligence Search layer.

3.3 AI Capability Discovery

The search layer identifies candidate AI systems based on their capabilities rather than their names alone. A candidate could be a general-purpose LLM, a domain-specific model, a tool-enabled agent, a retrieval system, or a verification service. Discovery may use metadata, historical performance, execution tests, reputation, and task-specific evidence.

3.4 AI Ranking and Selection

Candidate systems are ranked according to task fit. A simple conceptual scoring function can be written as:

$$Score(AI_i, T) = w_i C_i + w_q Q_i + w_r R_i + w_t T_i - w_k K_i$$

where C represents capability match, Q represents expected quality, R represents reliability, T represents time efficiency, and K represents cost. The weights depend on the task. This equation is not intended as a final algorithm; it illustrates that AI selection can be treated as a multi-objective retrieval problem.

3.5 AI-to-AI Communication

After selection, the system establishes communication among the chosen AI systems. Communication can include task decomposition, information exchange, intermediate results, questions, critiques, requests for verification, and negotiation over conflicting conclusions. In this framework, communication is not an isolated messaging layer. It is part of the search process because communication can reveal whether a selected AI is actually useful for the task.

This is the point at which the title of the paper becomes important. AI-to-AI Communication is the operational mechanism through which distributed intelligence can be accessed and combined.

3.6 Cross-AI Evaluation

Multiple AI outputs should not automatically be merged. One AI may be confident but wrong, while another may identify a critical limitation. Cross-AI evaluation therefore compares evidence, reasoning quality, consistency, tool results, and independent verification signals. Agreement can increase confidence, while disagreement can trigger additional search or re-selection.

3.7 Agreement-Conflict Analysis

The Agreement–Conflict Analysis block identifies agreements and contradictions among the responses generated by participating AI models. When significant conflicts or unresolved differences are detected, feedback is sent to the Intelligence Search block to adjust the

participating AI set as needed. The search process is then repeated with the updated set of AI models to improve response consistency and reliability. This feedback loop enables the proposed framework to adapt its search process based on the analysis results rather than following a strictly linear approach.

3.8 Intelligence Synthesis

The final stage combines validated contributions into a coherent result. Synthesis should preserve uncertainty where necessary rather than simply averaging opinions. The objective is not to maximize the number of participating AIs; it is to produce the best-supported answer for the task.

4. Discussion

The main argument of this paper is deliberately broader than a particular implementation. Search has historically been about locating useful information. Modern AI systems increasingly have access to diverse forms of computation and expertise. Once those systems can be discovered and contacted, the search problem can move from information retrieval toward intelligence retrieval.

AI-to-AI Communication is central to this transition because discovery alone is insufficient. Finding an appropriate AI does not guarantee a useful solution. The selected AI must be able to contribute, receive context, respond to other agents, defend or revise its result, and participate in a process that evaluates competing outputs.

One of the most important aspects to consider is the fundamental distinction between the proposed algorithm and Artificial General Intelligence (AGI). While AGI focuses on developing systems capable of performing a broad range of intellectual tasks [9], the proposed algorithm focuses on coordinating multiple AI models, evaluating their responses, and selecting the most appropriate answer. Rather than developing a single general-purpose intelligence, it aims to exploit the complementary capabilities of existing models through a structured intelligence retrieval framework. Thus, the proposed approach represents a complementary research direction, shifting the focus from individual model capabilities to collaborative intelligence, whose effectiveness requires empirical validation.

Alongside the various advantages of the proposed algorithm, it is important to recognize its potential limitations. Coordinating multiple AI models may increase computational costs and response latency, while access restrictions, API availability, and differences in communication protocols may complicate interaction between heterogeneous AI systems. Furthermore, variations in data formats, response structures, and model capabilities may require additional mechanisms for standardization and evaluation. Therefore, implementing the proposed framework effectively may depend on the availability of reliable communication interfaces, efficient coordination mechanisms, and scalable model-selection strategies.

Finally, the practical implementation of this idea could offer diverse benefits and applications, potentially transforming how scientific research and complex problem-solving are conducted. By coordinating multiple AI models and evaluating their outputs, the proposed framework could enhance intelligent search, facilitate scientific investigations, support complex decision-making, and improve technical troubleshooting. In particular, it could enable researchers to explore complementary perspectives generated by different AI systems, helping them investigate complex questions more systematically. However, realizing this potential will require practical implementation and empirical validation across different application domains.

5. Conclusion

This paper proposed AI-to-AI Communication as a central mechanism for a broader paradigm called Intelligence Search. The proposed perspective extends conventional search from finding information to finding, coordinating, evaluating, and synthesizing task-relevant AI capabilities.

The framework does not claim that multi-agent systems, agent discovery, model routing, multi-LLM collaboration, or AI-to-AI communication are individually new. Instead, it argues that these developments can be unified around a common research question: how can an AI system search for the intelligence it needs?

If this direction proves practical, future search systems may not simply return documents or generate answers. They may dynamically identify which AI systems should participate in solving a problem, allow those systems to communicate, test their outputs against one another, and synthesize the most reliable result. In this sense, the next generation of search may move from searching for information toward searching for intelligence.

References

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020. ([NeurIPS Papers](#))
- [2] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2024). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. COLM 2024.

- [3] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “ReAct: Synergizing Reasoning and Acting in Language Models,” *International Conference on Learning Representations (ICLR)*, 2023. ([arXiv](#))
- [4] Krasovitskiy, K., Christou, S., & Zeinalipour-Yazti, D. (2025). LLM-MS: A Multi-Model LLM Search Engine. IEEE 41st International Conference on Data Engineering Workshops (ICDEW 2025), 107–114. DOI: 10.1109/ICDEW67478.2025.00020.
- [5] Wu, B., Mammadli, A., Zhang, X., & Yilmaz, E. (2026). AgentSearchBench: A Benchmark for AI Agent Search in the Wild. arXiv:2604.22436.
- [6] Cui, Y., Chao, Y., & Du, C. (2026). AI Agent Discovery and Invocation Protocol. Internet-Draft draft-cui-ai-agent-discovery-invocation-02, July 2026.
- [7] Singh, A., Ehtesham, A., Kumar, S., Talaei Khoei, T., & Vasilakos, A. V. (2025/2026). Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. arXiv:2501.09136.
- [8] Yang, S., Li, Y., Lam, W., & Cheng, Y. (2026). Multi-LLM Collaborative Search for Complex Problem Solving. Findings of the Association for Computational Linguistics: ACL 2026, 42599–42614. DOI: 10.18653/v1/2026.findings-acl.2115.
- [9] M. R. Morris, J. Sohl-Dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, C. Farabet, and S. Legg, “Position: Levels of AGI for Operationalizing Progress on the Path to AGI,” in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, vol. 235, 2024, pp. 36308–36321.