Beyond Monoliths: Expert Orchestration for More Capable, Democratic, and Safe Language Models

Anonymous ACL submission

Abstract

This position paper argues that the prevailing trajectory toward ever larger, more expensive generalist foundation models controlled by a handful of companies limits innovation and constrains progress. We challenge this approach by advocating for an "Expert Orchestration" (EO) framework as a superior alternative that democratizes LLM advancement. Our proposed framework intelligently selects from many existing models based on query requirements and decomposition, focusing on identifying what models do well rather than how they work internally. Independent "judge" models assess various models' capabilities across dimensions that matter to users, while "router" systems direct queries to the most appropriate specialists within an approved set. This approach delivers superior performance by leveraging targeted expertise rather than forcing costly generalist models to address all user requirements. EO enhances transparency, control, alignment, performance, safety and democratic participation through intelligent model selection.

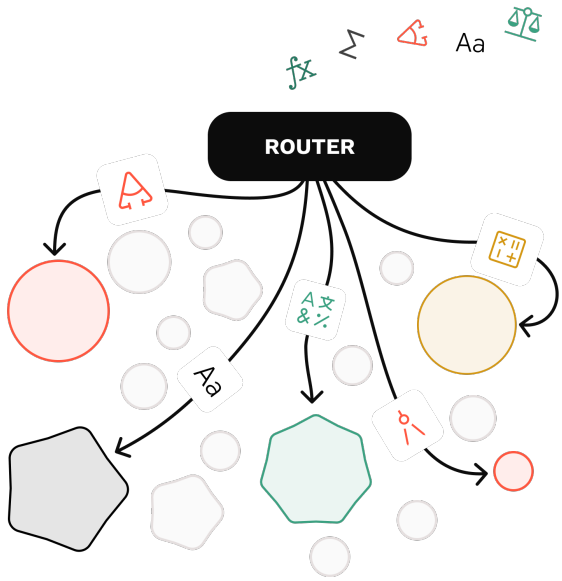

Figure 1: Expert Orchestration framework: The router leverages cached model capability analyses to categorize incoming prompts and dynamically route queries to the optimal model — whether Transformer-based or alternative architectures, large or small, specialist or generalist, cloud-based or edge-deployed — maximizing response quality while minimizing cost.

1 Introduction

The field of artificial intelligence (AI) has witnessed remarkable progress, largely driven by advancements in large language models (LLMs). Currently, users predominantly rely on monolithic frontier LLMs for queries and tasks. When these models fall short by producing hallucinations (Simhi et al., 2025; Zhang et al., 2023), showing bias (Gallegos et al., 2024), or lacking specialized knowledge (Kandpal et al., 2022) the typical response from both developers and users has been to attempt to "fix" these shortcomings through techniques like prompt engineering, RLHF, vector steering, or parameter-efficient fine-tuning. Unfortunately, task interference may cause "whack-a-mole" side effects compromising a large scale generalist model, as shown by researchers in studies for fine-tuning (Shumailov et al., 2024), steering (Stickland et al., 2024), RLHF (Kirk et al., 2023), and poor prompt selection (Cao et al., 2024).

We believe that it is fundamentally intractable (Varoquaux et al., 2025) to develop a single model capable of optimal performance across all possible tasks. Hence, this paper argues against the prevailing approach of building ever-larger generalist models. Patching individual models to perform well across all domains is akin to forcing a general practitioner to perform brain surgery rather than deferring the task to a neurosurgeon. Just as humans instinctively consult different experts based on their specific competencies, we should focus on identifying and leveraging the strengths of specialized models. Such specialization not only mirrors nat-

ural expertise distribution in human communities, but also offers a more effective path to addressing the limitations of current AI systems.

There are thousands of specialized models currently available on platforms like HuggingFace (Horwitz et al., 2025). By evaluating their strengths, we can leverage the diversity and specialization of these models. While both generalist and specialist models may require improvements, fixing specialist models is fundamentally more tractable for several reasons. First, specialist models operate in constrained domains with clearer evaluation metrics and more easily established ground truth. Second, the narrower input space dramatically reduces the testing matrix needed to ensure quality. Third, specialized human expertise can be more effectively applied to the limited domain. Finally, when specialists are improved, the benefits immediately propagate through the orchestration system without disrupting other domains, unlike monolithic models where fixes for one domain often cause regressions in others due to parameter interference (Shao and Feng, 2022; Saunders and DeNeefe, 2024; Xu et al., 2020). At the same time, Zaharia et al. (2024) argues state-of-the-art AI performance is increasingly driven not by scaling individual models, but by assembling compound systems composed of multiple coordinated components.

We posit that improvements in monolithic models aiming to handle all tasks is unsustainable. Rather, we propose a paradigm shift towards a framework we term *expert orchestration* (EO), comprised of specialized components: *Judges* that evaluate model capabilities across dimensions that matter to users (factuality, domain expertise, ethics, creativity, etc.), and *Routers* directing user queries to the most appropriate model(s) in a set of specialist & generalist models, based on user preferences. This approach improves control and monitoring, delivering superior answers at lower average cost creating a capable, democratic, and safe ecosystem.

Below, we outline limitations of the current landscape (Section 2), why interdisciplinary frameworks argue for change (Section 3), describe the EO framework (Section 4), argue it enhances LLM utility (Section 5), while acknowledging open research questions (Section 6) and alternative viewpoints exist (Section 7). Finally we urge adoption to secure a safer AI future (Section 8).

2 The Problem with Concentrating AI

Wu (2011) describes the recurring cycle, where information industries— such as radio and the internet— begin in a period of innovation but become consolidated by monopolies, which may suppress competition and innovation (Aghion et al., 2023). We highlight the growing risks of similar AI concentration, and call for scrutiny of democratic alternatives by the research community.

2.1 Market Dynamics

The economics of frontier AI development are increasingly shaped by powerful market incentives that favor “winner-take-all" outcomes, where a small number of dominant players capture the lion’s share of profits and influence. These companies operate with the expectation that the developers of the most capable generalist LLMs will capture the vast majority of the market, pursuing this consolidation with the potential to concentrate trillions of dollars and substantial deal-making power within a limited number of corporations.

This concentration raises concerns about equity and democratic governance, with potential systemic risks from such market dominance. International governments increasingly view AI concentration in a handful of US companies as a threat to sovereignty, local culture, and democratic values (LeCun, 2024)—a concern that extends beyond any single nation’s interests to the global distribution of AI capabilities.

EO addresses aspects of these market failures by lowering the resource threshold for meaningful contribution to the AI ecosystem and by creating a framework that naturally incorporates and highlights specialized excellence, regardless of the model creator’s scale or resources.

Current market dynamics create concerning misalignments regarding safety. Frontier companies, while often expressing commitment to safety, face real incentives to under-evaluate and under-report potential risks to expedite model releases. This rush to market is driven by intense competition and the desire to capture market share in the perceived “winner-take-all" dynamic. The DarkBench paper (Kran et al., 2025) demonstrates that models misrepresent their own capabilities and advantages over competitors, further complicating accurate risk assessment.

The underlying competitive dynamics often favor rapid capability advancement and market share

capture over meticulous safety assurance (Martian, 2025a). For companies selling access to increasingly powerful models, the motivation may be weak to dedicate significant resources to in-depth safety research that could slow capability advancements.

This concentration raises concerns about equity and democratic governance, with potential systemic risks from such market dominance.

EO addresses aspects of these market failures by lowering the resource threshold for meaningful contribution to the AI ecosystem and by creating a framework that naturally incorporates and highlights specialized excellence, regardless of the model creator's scale or resources.

2.2 Technical Challenges of monoliths

Limited User Insight into LLM "Thinking" Characteristics. Beyond the technical correctness, users are increasingly concerned with a range of underlying "thinking" characteristics. These include legality, morality, the absence of hallucinations, and the lack of gender or other biases. Currently, users have limited means to effectively communicate these criteria to LLMs and possess very limited ability to evaluate how well these models align with their desired thinking characteristics.

Frontier LLMs, however, often present themselves as being universally capable, without any clear differentiation regarding underlying thinking characteristics. While users can gain some limited control over these characteristics by their choice of LLM, and by employing specialized prompts, the actual impact and reliability of these methods remain uncertain.

This limitation is widely recognized in the alignment literature, where recent work emphasizes the importance of user-steerable LLMs and controllable generation. For instance, Bai et al. (2022) introduce Helpful and Harmless Assistant (HH-RLHF), where preferences are directly integrated into model behavior via human feedback loops and fine-tuning procedures.

Similarly, OpenAI's InstructGPT paper (Ouyang et al., 2022) shows that aligning LLMs with user intent through instruction-following dramatically improves user satisfaction and safety. However, these efforts are largely global alignment efforts so users do not have fine-grained, per-query control.

Users deserve more direct insight and specific control over the "thinking" characteristics of LLM behavior. None of the above methods match the explicit and modular control enabled by EO where each thinking characteristic (e.g., legality, bias, hallucinations) is explicitly evaluated and can be chosen by the user per query.

Monolithic Systems Are Less Controllable. While specialized models and frameworks that enable calling multiple models as tools do exist, ease of use considerations often lead most users to opt for a single LLM, with its inherent strengths and weaknesses, for all their queries.

This single LLM presents as a "monolith" that is sufficiently proficient across all query types. While this might hold true on average, it is demonstrably false at the individual query level. For many queries, other LLMs, potentially with specialist abilities directly relevant to the query, would be more suitable. Alternatively, a query might be simple enough (e.g., a basic arithmetic problem) that invoking a frontier model represents a wasteful expenditure of resources: money, time, and electricity.

The shift from monoliths to components also mirrors the move in NLP and CV towards modular sparse systems (Riquelme et al., 2021) and BASE Layers (Lewis et al., 2021), which show that task-specific experts outperform generalist models at lower cost and complexity.

The Modular Deep Learning paper (Pfeiffer et al., 2023) says "It remains unclear how to develop models that specialize towards multiple tasks without incurring negative interference and that generalize systematically to non-identically distributed tasks". The paper promotes modular deep learning as a potential partial solution to these challenges.

Monolithic Systems Are Cloud-based. The ever-increasing size of LLMs creates an inherent dependency on cloud infrastructure, ignoring the computational capabilities now present in edge devices where many user queries originate. Cloud-based LLMs rely on a stable network connection for inference, and the response time depends on network stability and speed. When inference occurs locally on edge devices, response time is significantly reduced, and applications can function with limited or no network connectivity (Tiwari, 2024). This cloud-centric approach adds unnecessary latency for simple queries, wastes electricity through redundant data transmission, and creates privacy concerns by requiring sensitive user data to travel to remote servers. Google AI Edge has demonstrated that on-device small language models can be effectively deployed on Android, iOS,

and Web platforms (Google AI Edge, 2025). These developments challenge the assumption that all AI inference must occur in the cloud.

EO naturally enables a hybrid approach where edge-device routers can evaluate whether user queries need escalation to cloud LLMs or can be handled by specialized SLMs installed locally. The deployment of SLMs on edge devices has emerged as a pivotal strategy to overcome cloud dependency, with organizations ranging from startups to tech giants recognizing this approach—NVIDIA argues that small language models represent the future of agentic AI, while successful deployments on devices like Raspberry Pi and Jetson Nano demonstrate that even resource-constrained hardware can achieve considerable performance improvements without compromising privacy or efficiency (Belcak et al., 2025; Premai Blog, 2025). Users could install specialist SLMs for topics they commonly query, allowing edge-device routers to handle routine requests locally while preserving cloud resources for complex tasks.

3 Specialization Works: Lessons from NLP and Other Fields

A growing body of research demonstrates that specialized large language and vision–language models consistently outperform their generalist counterparts when tasks demand domain expertise, structured reasoning, or cultural grounding. For example, domain-trained models such as MatSciBERT (Gupta et al., 2022), Me-LLaMA (Xie et al., 2024), and LawLLM (Shu et al., 2024) surpass larger foundation models on corpora in materials science, medicine, and law. Even in abstract reasoning domains, WizardMath and mechanistic analyses of arithmetic circuits (Luo et al., 2023; Quirke et al., 2025) show that compact, fine-tuned models can outperform broader architectures on specialized symbolic tasks. At the cultural and linguistic level, CultureLLM, LLM-jp, and AfriBERTa (Li et al., 2024; Aizawa et al., 2024; Ogueji et al., 2021) demonstrate that regional and multilingual adaptation improves coherence, fidelity, and social grounding across diverse societies.

Our position in EO builds directly on these findings: specialization is not an anomaly but a general principle observed across complex systems. Hayek's theory of distributed knowledge (Hayek, 1945) recognized that knowledge exists as "dispersed bits" across agents, favoring coordination over centralization. Smith's division of labor (Smith, 1776) and Ricardo's comparative advantage (Ricardo, 1817) formalized the same insight—overall efficiency increases when tasks are partitioned according to relative strengths, precisely what EO operationalizes through intelligent routing among models. Organizational and cognitive theories echo this view: Condorcet's Jury Theorem (Condorcet, 1785) and Hong and Page's diversity theorem (Hong and Page, 2004) show that diverse agents collectively outperform any single expert, while Fodor's modularity of mind (Fodor, 1983) and Minsky's society of mind (Minsky, 1986) describe intelligence as an emergent property of specialized, interacting modules. Even biological evolution reflects this pattern—the emergence of multicellular organisms with differentiated cell types marks a decisive leap in capability through the division of labor (Rüffler et al., 2012). Orchestration among specialized models similarly enhances collective intelligence: Park et al. (2023) showed that over a hundred LLM agents, each with distinct roles, collaboratively coordinate a complex social event more efficiently than a single monolithic model.

Taken together, these convergent results across domains affirm EO's premise: specialization and coordination are enduring principles of intelligent systems, not temporary engineering choices. (See App. A for further discussion.)

4 An Expert Orchestration Framework

The limitations of monolithic frontier LLMs call for alternative approaches. Here we outline the expert orchestration (EO) framework, a compelling vision designed to overcome several shortcomings.

The Role of Judges. At the core of EO are specialized models or systems called "judges" that objectively assess specific characteristics of LLM outputs. For instance, separate judges might evaluate factual accuracy, legal compliance, ethical adherence, or the presence of hallucinations and biases. While judges currently evaluate model responses to user queries, alternative approaches exist — Kadavath et al. (2022) demonstrate that LLMs can assess the validity of their own claims and predict which questions they can answer correctly.

Independent judges enable trust and transparency in EO. By concentrating on distinct evaluation dimensions, judges enable comprehensive assessment of LLM outputs across multiple critical

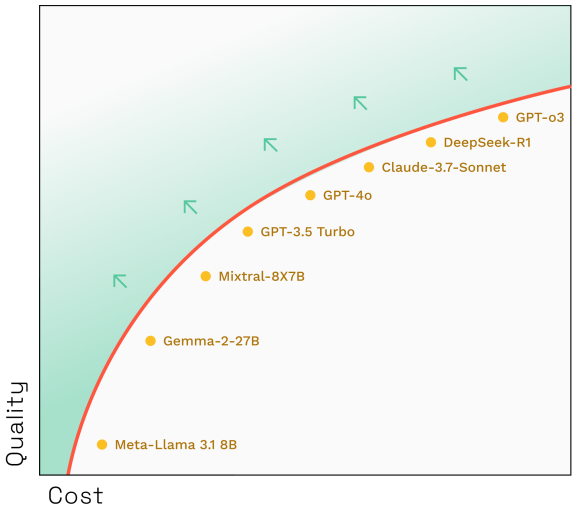

Figure 2: An experimental "meta-model" (red line) (Martian, 2025b) uses judges and routers to combine many models. It out-performs any single model (yellow points). More research is expected to move the Quality / Cost pareto curve "up and to the left" (green arrows).

characteristics.

The Role of Routers. The router receives queries and selects the optimal LLM(s) for each request (Prem Blog, 2025b). Routing decisions are informed by judge evaluations and user-specified preferences for response characteristics (Towards Data Science, 2025). For example, legal questions could be transparently routed to specialized legal models updated with recent case law, rather than relying on generalist models with uncertain legal reasoning (See Fig.1).

Routers also consider model specialization, operational cost, and response speed (Prem Blog, 2025a). A key advantage is their dynamic nature: EO readily adapts to new LLMs by incorporating them into the model set and using judges to assess their capabilities, enabling continuous evolution. Recent advances in cost-aware routing like HybridLLM and CARROT (Ding et al., 2024; Somerstep et al., 2025), plus adaptive MoE inference (Zhong et al., 2024) that dynamically selects experts based on task relevance and efficiency trade-offs, directly support EO implementation.

Superior Performance. Ensemble methods consistently outperform single models in machine learning (Hansen and Salamon, 1990; Dietterich, 2000; He et al., 2015; Devlin et al., 2018) (See Fig.2), and this holds for EO through both pre-hoc routing and post-hoc selection approaches.

Pre-hoc routing uses judges to train models that predict which LLM will perform best for each query. CARROT (Somerstep et al., 2025) predicts generation cost and quality from input prompts, achieving higher scores across cost ranges than any single model on RouterBENCH. Similarly, P2L (Frick et al., 2025), trained on LMSys ChatBot Arena preference data, topped the leaderboard by predicting user-preferred LLMs per query. Recent work (Shnitzer et al., 2024) even generalizes to unseen models at test time.

Post-hoc methods query multiple LLMs and select the best answer. While approaches like LLM-Blender (Jiang et al., 2023) are powerful, they incur high costs. More efficient cascading methods like FrugalGPT (Chen et al., 2023) query models sequentially, returning the first answer exceeding a quality threshold—achieving up to 2.4x cost reduction while maintaining quality. RouterBench experiments (Hu et al., 2024) confirm cascades outperform single models across cost levels, with performance heavily dependent on judge quality.

Integration of Specialized Models. EO enables seamless integration of specialized models into the broader ecosystem (Conclusion Intelligence, 2025). When specialized models demonstrate "best in class" performance in a domain, the router can prioritize them for relevant queries (Locaria, 2025). This removes the barrier requiring every innovator to match frontier LLM capabilities across all domains to gain market share. Instead, innovators can focus on achieving excellence within a narrower scope (Dredze et al., 2024), fundamentally democratizing model creation and fostering a vibrant community of specialist contributors.

While many specialized models will derive from foundation models, this strengthens our argument. EO realizes the full value of foundation models through selective specialization and strategic routing. Rather than forcing one model to perform optimally across all domains (an impossible task given parameter interference) orchestration leverages foundation capabilities while optimizing performance through specialized deployment. This represents a mature evolution analogous to how early integrated computer systems evolved into specialized components working in concert.

Model-Architecture Agnostic. EO has the advantage of being architecture-agnostic: Beyond transformers, EO can incorporate specialized models like CNNs for vision (Zhao et al., 2024), RNNs for streaming data (Mienye et al., 2024), GNNs for relational structures (Zhou et al., 2020), and diffusion models for generative tasks (Yang et al.,

2025). For instance, EO may route a graph reasoning task to a GNN, a time-series forecast to an RNN, and a textual explanation to an LLM, then combine outputs through judges. This capacity to integrate diverse architectures increases EO's utility, as different model architectures may excel in different modalities or constraints such as latency, energy efficiency, or symbolic reasoning.

5 How EO Enhances LLM Utility

EO offers substantial enhancements across several key dimensions of LLM utility, leading to a more robust, user-centric, and responsible ecosystem.

Increased Transparency and Trust. A significant benefit of EO lies in its inherent ability to increase transparency and build trust in LLM outputs (IBM, 2025). By employing dedicated, independent, and objective judges to evaluate specific characteristics of interest across a multitude of models, the framework provides users with a clearer understanding of the strengths and weaknesses of different LLMs in various domains (SmythOS, 2025; PMC, 2025).

EO parallels the rigorous, standardized evaluation practices found in safety-critical domains such as aviation, nuclear energy, and medical device manufacturing, where independent regulatory bodies and engineering frameworks are used to ensure system safety, reliability, and compliance (Leveson, 2016; Rushby, 1994; Storey, 1996). Organizations like the Vector Institute and DNV already provide independent evaluations of models and vendors, highlighting the importance of this objective assessment (GlobeNewswire, 2025; DNV Group, 2025).

Recent work shows the importance of exploring internal reasoning to improve public trust (Kook et al., 2022; Thakur et al., 2024).

Selection of Judges Empowers Users. The use of judges focusing on specific characteristics empowers users with greater control over responses (Phenx AI, 2025). EO allows a user to specify prioritized characteristics for individual requests, with the router directing queries to models best suited to provide aligned answers (Prem Blog, 2025a).

Decomposing Requests Improves Alignment, Control, and Accuracy. . EO facilitates decomposing complex requests into manageable steps, such as planning followed by execution phases (Eyelevel.ai, 2025). Specialized project models handle planning, with "supervisor" models reviewing results to enhance safety. Costing models estimate required resources, while execution steps are delegated to domain-specific models.

This decomposition provides natural monitoring points and reduces the "scope of control" of any single model, lessening reliance on potentially misaligned models and mitigating single points of failure. Untrustworthy models can be swapped out. Decomposition allows us to start developing robust control techniques now.

Realigning Market Incentives Towards Specialization and Competition. EO fundamentally restructures market dynamics by eliminating both the transaction costs and competitive moats that make specialized models economically unviable (Varoquaux et al., 2025). Currently, users face high switching costs when moving between different models for different tasks—learning new interfaces, managing multiple subscriptions, and remembering which model works best for what. These frictions make generalist models attractive despite inferior performance in specific domains. Simultaneously, incumbent companies build defensive moats by creating models that are "good enough" across many domains, making it hard for users to justify switching despite superior specialists existing. EO destroys both barriers: it removes transaction costs through seamless automatic routing behind a unified interface, while eliminating defensive moats by automatically choosing the best model for each task. This makes it impossible to defend market position through convenience rather than capability—companies must continuously earn their position through specialized excellence.

Organizations implementing EO face structural incentives that naturally promote ecosystem health. To maximize routing accuracy and customer value, they must maintain comprehensive, objective evaluations of available models on a per domain basis, combating the proliferation of contaminated benchmarks in training datasets (Dodge et al., 2021; Deng et al., 2023) for general capabilities. In doing so, they are also economically motivated to continuously seek out and integrate the most effective specialized models. This creates sustainable market demand for niche innovators while incentivizing transparency: orchestration providers gain credibility through verifiable model evaluations rather than capability hoarding, creating competitive pressure toward better measurement and disclosure.

Furthermore, the organization is naturally driven to publish objective "leaderboards" that rank mod-

els based on their performance across various capability areas. This transparency provides a clear benchmark for innovators, who then only need to create a model that excels in a specific area to gain recognition and potential integration into an EO implementation. EO stops the possibility of "best general model captures all value".

6 Research Directions and Challenges

EO, while promising, presents several key research questions that represent exciting areas for innovation. First, developing robust methodologies for evaluating models across diverse "thinking" characteristics beyond traditional metrics is essential, including bias (Team, 2025c), fairness (Team, 2025b), and hallucination detection (Team, 2025a).

Research is needed on utilizing multiple judges that reflect diverse user preferences to inform routing decisions. Deeply understanding model capabilities beyond simple benchmarking is necessary for optimal task matching.

Additional research directions include: (1) developing efficient and scalable routing algorithms that handle numerous models and complex preferences; (2) addressing the cold-start problem for new models with limited performance data; (3) exploring techniques for composing specialized models (Yang et al., 2024) to create more powerful capabilities; (4) studying broader ecosystem dynamics and impacts on competition and innovation; (5) applying dynamic model selection techniques (Brownlee, 2025) for adaptive routing; (6) developing theoretical models for when and how EO can improve performance such as with boosting; and (7) how to leverage ensemble methods (Chen et al., 2025) and cost-aware routing (Somerstep et al., 2025) to optimize performance and efficiency.

Research into different architectures for implementing judges – including fine-tuned specialized models, rule-based systems, and human evaluation integration – represents another critical area for investigation. Together, these research directions will help realize the full potential of EO while addressing its current limitations.

7 Relation to Existing Approaches

While EO proposes a shift in paradigm, it is important to situate this framework alongside existing strategies that have shaped the current AI landscape. We consider most of these as orthogonal to EO, and would still have their use cases.

Scaling Frontier LLMs. The dominant approach in large labs has been to continually scale generalist models. This delivers strong average performance but at rising economic and environmental costs, with diminishing returns at the margins. EO instead leverages frontier models only where necessary, while enabling smaller specialized models to contribute value, lowering barriers to entry.

Reinforcement Learning with Human Feedback (RLHF) and Fine-Tuning. While effective for specialized domains, they often introduce "whack-a-mole" regressions in generalist models where improvements in one area degrade others (Ouyang et al., 2022; Kirk et al., 2023). EO mitigates this brittleness by not requiring a single model to serve all purposes: weaknesses in an expert can be compensated by routing to another without destabilizing the system.

Agentic AI. Agentic frameworks expand the action space of a single LLM by enabling planning, tool use, and multi-turn reasoning (Singh et al., 2025). They increase capability but still rely on one model's internal judgment for when and how to act. EO instead governs *which* model should act, using independent judges and routers to select the most appropriate specialist. So EO can treat an agentic LLM as just another expert within its ecosystem, subject to routing and oversight.

Multi-Agentic AI. Recent work explores "societies" of LLM agents that collaborate via debate, role specialization, or simulation (Ye et al., 2025; Han et al., 2025). These systems demonstrate emergent capabilities, but typically rely on multiple instantiations of the *same* generalist model and depend on endogenous coordination. As a result, they face persistent challenges in task allocation, layered context and memory management, and ensuring reliable collective decisions (Han et al., 2025). EO reframes this paradigm as a *system-level meta-controller*: it *explicitly* curates heterogeneity by incorporating specialist models, assigns roles through routers, and enforces policy and quality constraints via judges. So EO subsumes some of the benefits of multi-agent interaction while adding verifiable oversight, auditability, and cost/latency controls—providing governance, transparency, and democratic participation.

Ensembling and Mixture-of-Experts (MoE). Approaches such as LLMBlender (Jiang et al., 2023), FrugalGPT(Chen et al., 2023), and MoE (Zhong et al., 2024) architectures combine multiple outputs or subnetworks to improve perfor-

mance. EO differs by orchestrating independent models across organizational boundaries, allowing new specialists to enter without retraining the whole system.

These perspectives highlight important avenues for ongoing reflection, implementation caution, and further research, which we believe strengthen rather than diminish the case for EO.

8 Conclusion: Towards a More Robust and Human-Aligned Future

The current dominance of monolithic frontier LLMs suffers from inherent limitations related to winner-take-all dynamics, misaligned safety incentives, barriers to entry for specialized models, limited user insight, and the inefficiencies of a one-size-fits-all approach.

EO offers a compelling alternative that addresses these shortcomings by introducing a framework composed of specialized evaluation models ("judges") and intelligent routing systems ("routers"). This approach promises higher quality answers at a lower average cost by strategically leveraging the strengths of diverse models, including both frontier and specialized ones.

The framework enhances transparency and trust through independent evaluation, empowers users with granular control over desired characteristics, improves alignment and accuracy through request decomposition, and fosters a more democratic and open ecosystem. Moreover, an organization implementing EO has incentives naturally aligned with safety and transparency.

If AGI does not emerge suddenly from a single generalist system, an EO framework could achieve AGI earlier than general models. Our approach enables the development now of strong safeguards that decrease potential extinction-level threats.

By addressing many limitations of the current paradigm and offering a path towards a more user-centric and responsible future, EO holds significant promise for shaping the next generation of language model applications.

9 Limitations

As with any emerging framework, EO invites thoughtful critique and warrants a balanced evaluation. Some limitations and alternative perspectives surfaced during the development of this work:

Centralization of Orchestration Infrastructure. Critics note that EO could shift gatekeeping from models to those controlling routers and judges, potentially recreating winner-take-all dynamics. However, unlike monolithic models, orchestration components are logically separable—enabling regulatory intervention, user customization (bringing own judges or selecting model families), and lower barriers to entry. The heterogeneity of routing needs across domains suggests market fragmentation rather than consolidation. Crucially, transparency requirements are far more enforceable for modular orchestration systems than for opaque monoliths, making democratic oversight tractable.

Corporate Incentives and Public Benefit Structures. This paper emphasizes misaligned incentives in frontier model development. Critics note that some organizations operate as or are transitioning to public benefit corporations (PBCs) and are legally permitted—and in some cases obligated—to prioritize societal welfare alongside shareholder value. This complicates a purely profit-motivated critique.

Latency and Cost Trade-offs. The orchestration of multiple models introduces questions about computational efficiency. Decomposing queries and routing them through specialized evaluators and responders may increase latency or system overhead in certain cases. These costs must be weighed against the benefits of specialized performance and may be mitigated through efficient pre-hoc routing strategies.

Applicability Beyond Language. Readers may view EO as specific to LLMs. In practice, the framework generalizes to other modalities—including vision, speech, and multimodal systems—where specialized components can also enhance performance, transparency, and control.

Sufficiency of Generalist Models. Critics may see current generalist models, particularly when augmented with tool use, as "good enough" for most applications. We believe this view underestimates both the current limitations and long-term risks. Specialized systems consistently outperform generalists in high-stakes or knowledge-intensive domains. EO offers structural benefits—transparent governance, distributed safety guarantees, and robust oversight—that generalist architectures and tool use alone cannot provide.

References

Philippe Aghion, Antonin Bergeaud, Timo Boppart, Peter J Klenow, and Huiyu Li. 2023. A theory of falling growth and rising rents. *Review of Economic Studies*, 90(6):2675–2702.

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, and 1 others. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small language models are the future of agentic ai. *Preprint*, arXiv:2506.02153.

Jason Brownlee. 2025. Dynamic classifier selection ensembles in python. *Machine Learning Mastery*.

Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. 2024. On the worst prompt performance of large language models. *arXiv preprint arXiv:2406.10248*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Dingqi Yang, Hailong Sun, and Philip S. Yu. 2025. Harnessing multiple large language models: A survey on llm ensemble. *arXiv preprint arXiv:2502.18036*.

Paul Christiano, Buck Shlegeris, and Dario Amodei. 2018. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.

Conclusion Intelligence. 2025. The rise of specialized language models (slms). https://conclusion.intelligence.com/. Accessed: 2025-05-21.

Marquis de Condorcet. 1785. *Essay on the application of analysis to the probability of majority decisions*. Paris: De l'imprimerie royale.

Robert A Dahl. 2008. *Democracy and its Critics*. Yale university press.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

John Dewey and Melvin L Rogers. 2012. *The public and its problems: An essay in political inquiry*. Penn State Press.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. *Multiple Classifier Systems*, pages 1–15.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subharata Mukherjee, Victor Ruhle, Laks VS Lakshmanan, and Ahmed Hassan Awadallah. 2024. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*.

DNV Group. 2025. Ai vendor capability assessment: Demonstrate trustworthiness of your ai solution. https://www.dnv.com/. Accessed: 2025-05-21.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Mark Dredze, Genta Indra Winata, Prabhanjan Kambadur, Shijie Wu, Ozan İrsoy, Steven Lu, Vadim Dabravolski, David Rosenberg, and Sebastian Gehrmann. 2024. Academics can contribute to domain-specialized language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5100–5110.

Eyelevel.ai. 2025. Optimizing rag systems with advanced llm routing techniques: A deep dive. https://eyelevel.ai/. Accessed: 2025-05-21.

Jerry A Fodor. 1983. *The modularity of mind: An essay on faculty psychology*. MIT press.

Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, Anastasios N Angelopoulos, and Ion Stoica. 2025. Prompt-to-leaderboard. *arXiv preprint arXiv:2502.14855*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

GlobeNewswire. 2025. Vector institute unveils comprehensive evaluation of leading models. https://www.globenewswire.com/. Accessed: 2025-05-21.

Google AI Edge. 2025. On-device small language models with multimodality, RAG, and function calling. Google Developers Blog. Accessed: 2025-05-20.

Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. 2022. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Jürgen Habermas. 2015. *Between facts and norms: Contributions to a discourse theory of law and democracy*. John Wiley & Sons.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, and Zhaozhuo Xu. 2025. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*.

Lars Kai Hansen and Peter Salamon. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001.

Friedrich A Hayek. 1945. The use of knowledge in society. *The American Economic Review*, 35(4):519–530.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389.

Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. 2025. Charting and navigating hugging face's model atlas. *arXiv preprint arXiv:2503.10633*.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. 2024. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031*.

IBM. 2025. What is explainable ai (xai)? https://www.ibm.com/. Accessed: 2025-05-21.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. Ai safety via debate. In *International Conference on Learning Representations Workshop*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nelson Schiefer, Zac Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. *arXiv preprint arXiv:2211.08411*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Lucas Kook, Andrea Götschi, Philipp FM Baumann, Torsten Hothorn, and Beate Sick. 2022. Deep interpretable ensembles. *arXiv preprint arXiv:2205.12729*.

Esben Kran, Jord Nguyen, Akash Kundu, Sami Jawhar, Jinsuk Park, and Mateusz Jurewicz. 2025. Darkbench: Benchmarking dark patterns in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR.

Yann LeCun. 2024. Interview with lex fridman. LessWrong.

Nancy G Leveson. 2016. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. 2021. Base layers: Simplifying training of large, sparse models. *International Conference on Machine Learning*, pages 6265–6274.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Locaria. 2025. Specialized models rise as llms stall. https://locaria.com/. Accessed: 2025-05-21.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

James Madison. 1788. The federalist no. 51: The structure of the government must furnish the proper checks and balances between the different departments. *Independent Journal*, 6.

Martian. 2025a. Safety vs capitalism - blog. https://blog.withmartian.com/post/ai-safety-incentives/. Accessed: 2025-05-21.

Martian. 2025b. Up and to the left! how martian uses routing to push the pareto frontier. Martian Blog. Accessed: 2025-10-03.

Ibomoiye Domor Mienye, Theo G. Swart, and George Obaido. 2024. Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, 15(9).

Marvin Minsky. 1986. *The society of mind*. Simon & Schuster.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st workshop on multilingual representation learning*, pages 116–126.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. 2023. Modular deep learning. *arXiv preprint arXiv:2302.11529*.

Phenx AI. 2025. The art of traffic control: Mastering llm routing. https://phenx.ai/. Accessed: 2025-05-21.

PMC. 2025. Editorial: Explainable ai in natural language processing. https://www.ncbi.nlm.nih.gov/pmc/. Accessed: 2025-05-21.

Prem Blog. 2025a. Balancing llm costs and performance: A guide to smart deployment. https://premai.io/blog/. Accessed: 2025-05-21.

Prem Blog. 2025b. Llm routing: Costs optimisation without sacrificing quality. https://premai.io/blog/. Accessed: 2025-05-21.

Premai Blog. 2025. Small language models (SLMs) for efficient edge deployment. Blog post. Accessed: 2025-03-04.

Philip Quirke, Clement Neo, and Fazl Barez. 2025. Understanding addition and subtraction in transformers.

David Ricardo. 1817. *On the principles of political economy and taxation*. John Murray, London.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595.

Claus Rüffler, Joachim Hermisson, and Günter P Wagner. 2012. Evolution of functional specialization and division of labor. *Proceedings of the National Academy of Sciences*, 109(6):E326–E335.

John Rushby. 1994. Critical system properties: Survey and taxonomy. *Reliability Engineering & System Safety*, 43(2):189–219.

Danielle Saunders and Steve DeNeefe. 2024. Domain adapted machine translation: What does catastrophic forgetting forget and why? *arXiv preprint arXiv:2412.17537*.

Chenze Shao and Yang Feng. 2022. Overcoming catastrophic forgetting beyond continual learning: Balanced training for neural machine translation. *arXiv preprint arXiv:2203.03910*.

Tal Shnitzer, Zichang Lin, Cheng Yang, Hao Cheng, Shelby Begnaud, Tianyi Zhang, Heng-Tze Cheng, Soham Chatterjee, and Hongyang R Jin. 2024. Universal model router improves llm performance across diverse benchmarks. *arXiv preprint arXiv:2502.08773*.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. pages 4882–4889.

Ilia Shumailov, Zakhar Shumailov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. Trust me, i'm wrong: High-certainty hallucinations in llms. *arXiv preprint arXiv:2502.12964*.

Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. 2025. Agentic reasoning and tool integration for llms via reinforcement learning. *arXiv preprint arXiv:2505.01441*.

Adam Smith. 1776. *An inquiry into the nature and causes of the wealth of nations*. W. Strahan and T. Cadell, London.

SmythOS. 2025. Explainable ai in natural language processing: Enhancing transparency and trust in language models. https://smythos.com/. Accessed: 2025-05-21.

Seamus Somerstep, Felipe Maia Polo, Allysson Flavio Melo de Oliveira, Pratyush Mangal, Mirian Silva, Onkar Bhardwaj, Mikhail Yurochkin, and Subha Maity. 2025. Carrot: A cost aware rate optimal router. In *ICLR 2025 Workshop on Foundation Models in the Wild*. https://huggingface.co/CARROT-LLM-Routing.

Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R Bowman. 2024. Steering without side effects: Improving post-deployment control of language models. *arXiv preprint arXiv:2406.15518*.

Neil Storey. 1996. *Safety-Critical Computer Systems*. Addison-Wesley, Harlow, England.

Aisera Research Team. 2025a. Llm evaluation: Key metrics, best practices and frameworks. Technical report, Aisera.

Forbes Councils Technology Team. 2025b. Ai & fairness metrics: Understanding & eliminating bias. *Forbes Councils*.

Test.io Research Team. 2025c. Llm bias: Understanding, mitigating and testing the bias in large language models. Technical report, Test.io.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Pankaj Tiwari. 2024. Edge AI: Deploying large language models for smarter devices. Medium: Accredian. Accessed: 2024-09-13.

Towards Data Science. 2025. Llm routing—intuitively and exhaustively explained. https://towardsdatascience.com/. Accessed: 2025-05-21.

Gaël Varoquaux, Alexandra Sasha Luccioni, and Meredith Whittaker. 2025. Hype, sustainability, and the price of the bigger-is-better paradigm in ai. *arXiv preprint arXiv:2409.14160*.

Michael Walzer. 2008. *Spheres of justice: A defense of pluralism and equality*. Basic books.

Tim Wu. 2011. *The master switch: The rise and fall of information empires*. Vintage.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, and 1 others. 2024. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3.

Ying Xu, Xu Zhong, Antonio Jose Jimeno Yepes, and Jey Han Lau. 2020. Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2025. Diffusion models: A comprehensive survey of methods and applications. *Preprint*, arXiv:2209.00796.

Rui Ye, Shuo Tang, Rui Ge, Yaxin Du, Zhenfei Yin, Siheng Chen, and Jing Shao. 2025. Mas-gpt: Training llms to build llm-based multi-agent systems. In *International Conference on Machine Learning (ICML)*. Poster.

Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, and Ali Ghodsi. 2024. The shift from models to compound ai systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. 2024. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4):99.

Shuzhang Zhong, Ling Liang, Yuan Wang, Runsheng Wang, Ru Huang, and Meng Li. 2024. Adapmoe: Adaptive sensitivity-based expert gating and management for efficient moe inference. *arXiv preprint arXiv:2408.10284*.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.

A Appendix A: Theoretical Foundations for Expert Orchestration

This appendix provides detailed theoretical support for the Expert Orchestration framework from multiple disciplines.

A.1 Economic Theories of Distributed Knowledge and Market Structure.

Friedrich Hayek's seminal work on distributed knowledge (Hayek, 1945) provides a powerful economic framework supporting EO. Hayek argued that knowledge in society exists as "dispersed bits of incomplete and frequently contradictory knowledge which all the separate individuals possess," never in "concentrated or integrated form" in any single mind. This impossibility of centralizing all knowledge leads to the superiority of market mechanisms over central planning: markets function as information processors that coordinate distributed expertise through price signals.

Adam Smith's theory of the division of labor (Smith, 1776) illustrates how breaking complex tasks into specialized functions dramatically increases productivity. Smith further observed that "the division of labor is limited by the extent of the market", meaning specialization increases as markets grow. This principle applies directly to models:

as the demand for capabilities expands, we should expect greater specialization of models rather than continued focus on general-purpose systems.

David Ricardo's theory of comparative advantage (Ricardo, 1817) extends this insight, showing that even when one agent is superior at all tasks, the total output is maximized if agents specialize according to their relative strengths.

A.2 Organizational Theory: Collective Decision-Making and Diversity.

Condorcet's Jury Theorem (Condorcet, 1785) provides mathematical proof that groups of independent decision-makers with better-than-random accuracy consistently outperform individuals, with reliability approaching certainty as group size grows. This applies directly to EO, where specialized judge models serve as an "expert jury" providing more reliable assessment than any single generalist model.

Lu Hong and Scott Page's diversity theorem (Hong and Page, 2004) extends this insight, proving that "groups of diverse problem solvers can outperform groups of high-ability problem solvers." EO leverages this principle by maintaining diverse specialized models, each bringing distinct problem-solving approaches to user queries.

A.3 Cognitive Science: Modularity and Distributed Intelligence.

Cognitive science provides compelling evidence that intelligence naturally emerges from specialized, interacting components rather than monolithic processors. Jerry Fodor's "Modularity of Mind" theory (Fodor, 1983) demonstrates that human cognition comprises domain-specific modules specialized for particular functions like language or vision, each operating with some independence from others. This modularity enables both efficiency and robustness—when one module fails, others continue functioning.

Building on this foundation, Marvin Minsky's "Society of Mind" theory (Minsky, 1986) offers a direct parallel to EO. Minsky proposed that intelligence emerges from "the interaction of many small, simple parts" without requiring a complex central controller: "a model of the human mind more like a democracy than a supercomputer." Recent AI research has validated this approach: Park et al. (2023) demonstrated that over a hundred specialized LLM agents working together can outperform any single model on complex tasks by collaborating and sharing information.

A.4 Biological and Evolutionary Frameworks.

The evolution of multicellular life provides a compelling analogy for EO. Single-celled organisms function as generalists, handling all life processes internally. The transition to multicellularity involved cells specializing into different types (muscle, nerve, blood, etc.), dramatically increasing the organism's capabilities. As Rüffler et al. note, "division of labor among functionally specialized modules occurs at all levels of biological organization" and represents a major evolutionary trend because specialization enables higher performance Rüffler et al. (2012).

A.5 Routing as a Form of Democratic Algorithmic Institution.

EO reflects key democratic values: participation, accountability, and distributed influence. Robert Dahl emphasizes that democracy depends on broad inclusion and equal ability to shape outcomes (Dahl, 2008), while Jürgen Habermas underscores the role of open, reasoned dialogue in legitimizing decisions (Habermas, 2015). John Dewey sees democracy as collective problem-solving rooted in everyday association (Dewey and Rogers, 2012).

EO echoes these ideals by lowering barriers for niche model creators, enabling a wider range of contributors to offer specialized capabilities. Through open evaluation and fair task routing, it promotes meaningful participation and healthy competition. Like the U.S. system of checks and balances, this distribution of influence helps prevent dominance by any single actor, fosters fairness, and supports systemic stability ((Madison, 1788). This pluralistic structure enables excellence across diverse domains and interests—an ideal at the heart of Walzer's argument for justice through distinct but coexisting spheres of merit (Walzer, 2008).

A.6 Alignment and Safety Approaches.

The safety via debate framework (Irving et al., 2018) proposes training agents to engage in adversarial debates about questions, with a human or judge model determining which agent provides the most convincing answer. This approach uses multiple systems with potentially opposed viewpoints to surface flaws in each other's reasoning, improving the trustworthiness of answers. EO naturally incorporates this debate-like structure through its judge models.

Christiano et al. (2018)'s Iterated Distillation and Amplification (IDA) alignment framework parallels EO principles. IDA starts with humans or simple models breaking complex tasks into smaller sub-questions, answering those questions, and then

aggregating the answers. This decomposition approach is then distilled into a more efficient model, which is iteratively amplified through additional decomposition.

A.7 Synthesis.

Across economics, cognition, biology, and organizational theory, specialized coordinated systems consistently outperform monolithic designs for complex tasks. From Hayek's distributed knowledge to Minsky's society of mind to multicellular evolution, the pattern is clear: complex capabilities emerge through orchestrated interaction of specialized components, not through scaling generalist systems. EO applies these proven principles to AI, creating systems that are more capable, transparent, and democratically governable than monolithic alternatives.