
“Qualis Apes Aestate”: Forecasting Multi-Agent Systems*

Michael Mulet
Independent

Jonas Müller
Technische Universität
Berlin

Nanubala Gnana Sai
Shipsy

Charles
Renshaw-Whitman
Independent

Moritz Weckbecker
Fraunhofer HHI

With
Apart Research

Abstract

The agglomeration of AI agents into multi-agent systems may play an important role in the development of frontier AI capabilities, as well as in the economics of AI development and deployment. We discuss the possibility of a differential acceleration of multi-agent systems’ capabilities. We present a comparative benchmarking of a variety of single- and multi-agent systems on the ADeLe dataset. It is found that multi-agent systems do not generally confer increased accuracy on the benchmark considered.

Keywords: Multi-Agent, Forecasting, Benchmarking

Acknowledgments: We thank **EleutherAI** for donated compute and **Lambda Labs** for cloud credits. We extend our further gratitude to Ronak Mehta, Paulo Bova, and Sami Jawhar for their advice. J.H. Orallo for his encouragement.

Code/Artifacts: <https://github.com/mmulet/ai-forecasting-hackaton-2025>

*Research conducted at the AI Forecasting hackathon
<https://apartresearch.com/sprints/the-ai-forecasting-hackathon-2025-10-31-to-2025-11-02>, 2025

Section 1: Context and Core Research Question

The METR timelines paper[1] paints a very vivid trend of increasing AI capabilities — we wish to examine the potential effects of multi-agent agglomerations (multi-agent systems, MAS) on the economic and intellectual characteristics of frontier AI. Do MAS perform better, and if so, how and when? We are interested because a rise in MAS systems may impact timelines in two distinct ways: first, if MAS systems are, or become, more capable than single-agent systems (SAS), they can directly push forward the capabilities frontier, and second, MAS systems’ capabilities may depend on agents’ individual capabilities in a qualitatively distinct manner from SAS — it is conceivable that slight increases in model capability might *differentially* accelerate MAS capabilities, e.g., by improved communication, other-agent-modelling, in-context adaptability, etc. This applies both to base agent model intelligence, as well as to improved infrastructure for facilitating AI agents. We think the possibility (likely or not) of such a step-change warrants an even-handed examination of MAS compared to SAS, along with a thoughtful examination of how MAS’ present failures may be differentially amenable compared to SAS.

Due to length limitations, we do not present here a detailed analysis of the characteristics of MAS a propos timelines; **please see appendices B and C.** We’re rather proud of these and regret the omission.

Section 2: Experimental Determination of MAS Capabilities

To assess MAS capabilities, we design an experiment comparing single-agent and multi-agent systems on tasks from a benchmarking dataset.

Experiment Design Criteria

Tasks: To effectively test system capabilities, evaluations should span multiple subject areas with tasks ordered by human-perceived difficulty or solving time. Critically, tasks must not saturate, i.e. models should still fail on the hardest/longest tasks, and should be automatically verifiable.

Models: To determine whether and for which models multi-agent systems improve capabilities, evaluations should test various models differing in size, family, and release date.

Experimental Procedure

Tasks: To implement our experiments we followed the methodology used in Zhou et al [2] and used a subset of the Multiple Choice questions in ADeLe v1.0 Battery [2] which includes 63 tasks from 20 benchmarks, each of the samples in ADeLe are annotated 18 rubrics for Demand-Level-Annotation (DeLeAn v1.0) of general scales to allow the ability to examine ability profiles of LLMs, comprehensively

explaining what they can and cannot do. For example the Attention and Scan dimension measures the ability to “Focus on or locate specific elements within a given stream of information or environment in the whole process of solving a task.”. Each rubric-task pair is assigned a demand level from 0 to 4, indicating how much that dimension is required to complete the task. As in, [2] we create a “lite” subset of the ADeLE dataset for faster processing, using a subset of 2000 sample questions total. In the future, we would like to run on the full dataset. We use the Inspect framework [4] integrated to evaluate our models on the dataset.

Models: We test Qwen2.5-3B, Mistral-7B-v0.2, SmolLM3-3B, Qwen2.5-14B, SmolLM2-360M, SmolLM2-1.7B, LLama-3-3B, LLama-3-8B. The zoo of models ensures variety across model families and scales.

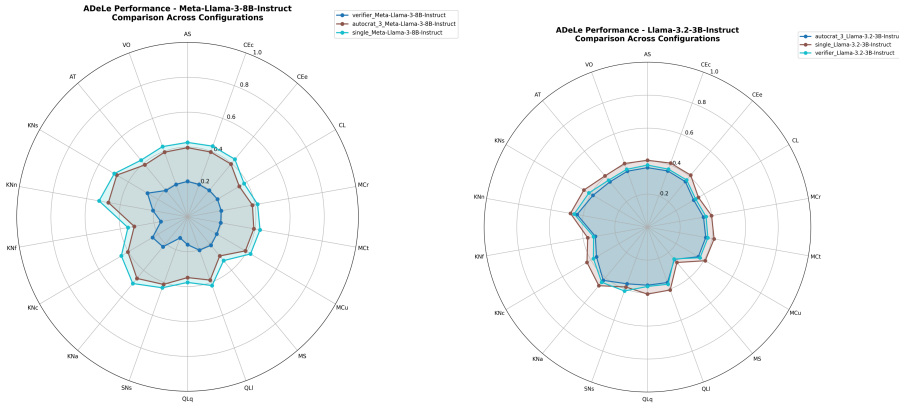
MAS setups: For each model, we compare a SAS with two different MAS with varying role assignments. The concrete instructions given to the agents can be found in Appendix A.2. The two MAS configurations we use are:

Guesser-Verifier: A two agent system, in which one agent is tasked with answering the prompt, while the other agent verifies the proposed solution. If judged adequate, the solution is given as output. If not the task is sent back to the prior agent with feedback on what to improve.

Autocrat: A four agent system, in which three agents are tasked with answering the prompt independently, and one autocratic agent that decides what solution is deemed to be the best response. This solution is given as output.

Metrics: We evaluate each model-system combination on the multiple choice subset and measure accuracy for each of the 18 annotated rubrics, where a task is considered to test a rubric if its demand level is at least 1.

Section 3: Experiment Results and Discussion



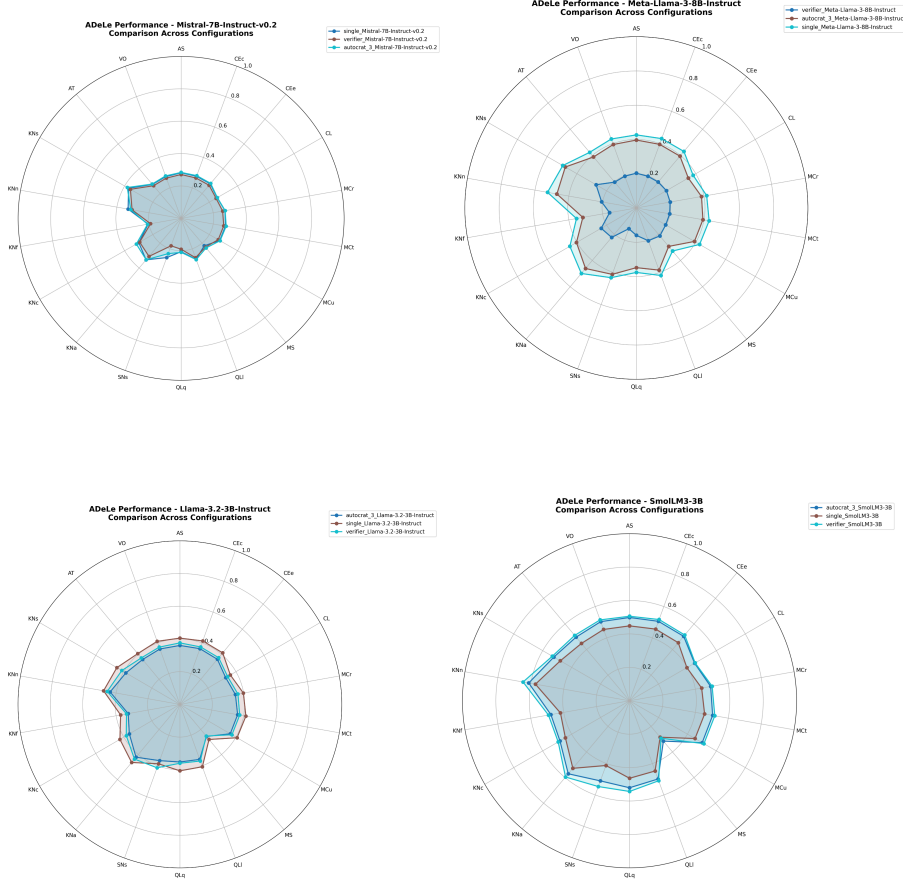


Figure 1. Adele Performance per model vs agent. For each model we ran, we evaluated its attributes and compared how it performed in the Single Agent, Autocrat, and Verifier configurations. For results on Qwen, check Appendix E, Figure 2. For analysis vs. demand level, see Appendix E, Figure 3.

Figure 1 demonstrates that, in most configurations, only one MAS case realized an improvement over the SAS baseline: SMOLM-3B showed consistent all-around improvement in the MAS, the largest difference being in SNs (spatial reasoning). This is largely a negative result, it shows that the application of MAS does not positively affect performance for the cases under consideration. In the MACNET[3] paper, they investigate how the performance of MAS scales as the number of agents increases (tested over many different topologies). They found that at 3-4 agents (the number of agents tested in this paper) the performance of MAS remain about the same as SAS. They only see a significant performance increase as the number of agents increases to about 15-25. This result matches ours, but unfortunately, the increased scale was above our compute and time budget for this weekend hackathon. It is an interesting avenue for future work.

Implications

Considering the results of our experiments (the limited scope of which notwithstanding) along with the analysis presented in the appendices, we estimate that strongly multi-agent systems are unlikely ($<15\%$) to catalyze a major

increase in capabilities within the coming year. If such comes to pass, we expect increased effort to go towards developing frontier models from more diverse architectural bases. We expect that economic pressures will continue to encourage model miniaturization, and that some applications like agentic coding will profit by dividing labor between intelligent overseer agents and cheaper worker agents.

Directions for Future Work

We see substantial potential in extending this research and are excited to explore several promising directions for future work.

Extend datasets Given more time, extending our experiments to the full ADeLe dataset and other capability benchmarks such as [1] would be valuable. We hypothesize that MAS systems might show greater improvements on [1], as our evaluation focused on multiple choice questions while [1] contains long coding tasks that could better leverage MAS advantages.

Extend to SOTA models: To enhance the quantitative evaluation and gather more robust data for forecasting, testing these experiments on state-of-the-art models would be valuable but was infeasible during the hackathon due to time, resource, and accessibility constraints.

Investigate different MAS setups: Different MAS setups (such as specialised experts) might increase capabilities for different tasks.

Scale number of agents: Prior work [3] suggests that capabilities of MAS systems depend on the number of deployed agents. It would be therefore interesting to see how the performance of different MAS setups (e.g. Autocrat) scale with the number of agents.

Compare resource consumption: While we investigate systems without time or resource constraints, examining capabilities relative to token usage, wall time, API costs, and compute requirements would provide valuable insights into their applicability in real-world scenarios.

Combine different models: Currently our approach uses the same model for every agent in the MAS, future work could see whether combining different model sizes or families improves MAS frameworks.

References:

[1] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, & Lawrence Chan. (2025). Measuring AI Ability to Complete Long Tasks.

[2] Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M. Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E. Prunty, Zongqian Li, Pablo Sánchez-García, Kexin Jiang Chen,

Pablo A. M. Casares, Jiyun Zu, John Burden, Behzad Mehrbakhsh, David Stillwell, Manuel Cebrian, Jindong Wang, Peter Henderson, Sherry Tongshuang Wu, Patrick C. Kyllonen, Lucy Cheke, Xing Xie, & José Hernández-Orallo. (2025). General Scales Unlock AI Evaluation with Explanatory and Predictive Power.

[3] Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, & Maosong Sun. (2025). Scaling Large Language Model-based Multi-Agent Collaboration.

[4] AI Security Institute, UK. 2024. “Inspect AI: Framework for Large Language Model Evaluations.” https://github.com/UKGovernmentBEIS/inspect_ai.

Appendix A: Security Considerations

In this Appendix, we will outline potential limitations of our approach and possible directions for further research to both address these limitations and enhance our findings via different directions.

Appendix A.1: Potential Limitations

Due to time constraints, we limited our experiments to the Multiple choice (MC) questions of ADeLe which limits our approach as we do not benchmark on a diverse set of problems. Additionally, we subsample 2000 of the 10000 MC questions to speed up evaluation, which gives us a relatively low sample size for some attributes and demand levels. For example, as seen in Figure 15 we have no sample for demand level 4 or 5 for the VO attribute.

Additionally, we noticed that some of the MC questions come in a different format. For example:

Let $\triangle ABC$ have incenter I and circumcenter O with $\overline{IA} \perp \overline{OI}$, circumradius 13 , and inradius 6 . Find $AB \cdot AC$. Please think step by step, and then display the answer at the very end of your response. The answer is an integer consisting of exactly 3 digits (including leading zeros), ranging from 000 to 999, inclusive. For example, the answer might be 068 or 972. If you cannot determine the correct answer, take your best guess. Remember to have the three digits as the last part of the response.
Let's think step by step before providing your final answer. However, you must conclude your assessment with the statement: 'Thus, the correct answer is: X', where 'X' is simply the letter corresponding to the answer option you've determined to be correct.

With metadata like this:

```
{'answer_format': 'MC', 'answer': '468'}
```

None of the models answered this correctly due to us only filtering for the format 'Thus, the correct answer is: X' when checking whether or not the answer is correct.

This format, we never encountered in any other MC question we investigated. The standard looks like

```
{'answer_format': 'MC', 'answer': 'H. suffers more than the victim.'}.
```

We are not sure whether this is intended or there might be a mislabeled datapoint.

Appendix A.2: Agent Instructions

```
head_agent: Agent

agents:

  - name: Agent

    instructions: You are a problem-solving agent. Analyze the task and
provide your best solution.
```

Figure 7.5: The instructions for a single Agent

```
# Architecture: 3 identical agents feed solutions to 1 autocratic
decision-maker

head_agent: Decision Maker

agents:

  # Agent 1

  - name: Agent 1

    instructions: You are a problem-solving agent. Analyze the task and
provide your best solution.

  # Agent 2

  - name: Agent 2

    instructions: You are a problem-solving agent. Analyze the task and
provide your best solution.
```

```

# Agent 3

- name: Agent 3

  instructions: You are a problem-solving agent. Analyze the task and
provide your best solution.


# Autocratic Decision Maker - Receives all solutions and makes final
decision

- name: Decision Maker

  instructions: |

    You are the autocratic decision maker. You will give the task to
all 3 models you can hand off too and receive their solutions.


    Your role is to:

    - Review all 3 proposed solutions

    - Evaluate their quality and correctness

    - Make the FINAL authoritative decision on which solution is best

    - Reproduce this solution as the final answer


    You have ultimate authority. Select the best solution.

  handoffs:

    - Agent 1

    - Agent 2

    - Agent 3

```

Figure π : The instructions for a AutoCrat configuration

```

# Architecture: Guesser and Verifier

```

```

head_agent: Verifier

agents:

  # Guesser

  - name: Guesser

    instructions: |

      You are a problem-solving agent. Analyze the task and provide your
      best solution.

      Your solution is judged by a different LLM. If it thinks your
      solution is not good enough,

      you will be given the same task again together with feedback.

      In this case, provide a different solution.

  # Verifier

  - name: Verifier

    instructions: |

      You are a judge agent. Your role is to:

      - Hand off the task to the Guesser agent

      - Evaluate the Guesser's solution for correctness and quality

      - If the solution is good enough, accept it and provide it as the
      final answer

      - If the solution is not good enough, explain why and hand back to
      Guesser Agent with feedback for another attempt

      Be critical but fair in your evaluation. You have ultimate authority.

```

```
handoffs:  
  
- Guesser
```

Figure ∞ : Instructions for the Guesser-Verifier Configuration

AI/LLM prompts used in your project:

Despite being over 15 pages long, we actually wrote this report ourselves (Charles really likes to write, and, as we have learned, loves Latin)

Appendix B: Analysis of MAS for timelines

Possible Mechanisms of Differential MAS Capabilities Acceleration

The central question is if the unique characteristics of MAS will offer a differential acceleration over that of the capabilities of their constituent models. A few possible mechanisms for such an acceleration are:

- **Reliability/robustness differences:** MAS requires more scaffolding design — presently ‘finicky’ — but offers unique ways to ameliorate problems (e.g., checks-and-balances, or new-agent instantiation to enhance ‘cognitive flexibility’).
- **Parallelizability & adaptive wall-clock speed:** parallel agents; dispatch smaller models for simpler tasks.
- **Improved context management:** inter-agent communications are necessarily summary in nature.
- **Improved modularization:** permits more comprehensive optimization; but non-local failures may be harder to understand.
- **Off-paradigm agents:** Mamba, diffusion LLMs, EBMs, and distinct modalities (image editors/generators) likely have very different capability profiles from frontier LLMs; even failed paradigm shifts and non-frontier developments may still advance MAS capabilities.
- **Deployment scalability:** MAS is the only way to scale deployment of many AIs to a single, sufficiently important task (e.g., research); more agent infrastructure will facilitate “not stepping on each others’ toes,” and inter-agent coordination may become non-optional.
- **Resource Pareto frontier:** additional design variables mean a Pareto frontier for allocating inference GPUs.
- **Distillation & miniaturization:** as inference costs decline precipitously, there will open up opportunities to profit by the deployment of many cheap agentic intelligences - this will only add to the economic pressures driving labs’ rush to miniaturize their models

Reasons to Think MAS Might Be More Useful Than SAS

We wish to characterise the main qualitative strengths afforded by the construction of multi-agent systems over and above their constituents. The most apposite of these are:

- **Speed on inherently parallelizable tasks:** MAS may dispatch numerous agents to rapidly solve parallel components of any task, decreasing wall-clock time to completion
- **Larger effective context size:** Having more agents entails a system-wide increase in context size, provided it can be utilized effectively.
- **Context conservation:** Agents can preserve their context usage by delegating tasks to smaller models and consuming only a summary of their outputs.

- **Specialization of concerns:** Agents may be assigned distinct roles within a collective, allowing individual optimization of different sub-components
- **Checks and balances:** Separate model instances may enforce standards more rigorously than a conflicted single agent, allowing for more robust detection and resolution of uncertainty.
- **Diversity dividend:** Agents with different capabilities profiles may, in effect, combine the best each model has to offer.
- **Flexible cognition:** Different model instances may more neutrally evaluate the merits of continuing to pursue a given approach to a problem - ‘Science progresses one funeral at a time’
- **Increased clarity from verbalization:** The necessity of inter-agent communication may force models to form more cogent explanations of their actions
- **Adaptable resource consumption:** Smaller models may be dispatched for simpler tasks, or modulated according to instantaneous inference costs.

Reasons to Think SAS May Be Better Than MAS

Having discussed potential advantages of MAS, it bears emphasising that, borderline cases like GPT5’s router design notwithstanding, frontier models are not MAS, nor even are most deployments. Some reasons supposed gains have failed to actualize:

- **Communication overhead:** information may need to be communicated multiple times between various agents, and further effort expended in coordination of their activities
- **Lost in translation:** Agents may miscommunicate, convey incorrect/ambiguous/inapposite info, perhaps due to poor ‘other-modeling’ of what to hand off.
- **Adversariality:** Agents meant to “check” one another may cause deadlock by being too exacting.
- **Decentralization of knowledge:** Agents neither have shared knowledge, nor, by default, knowledge of other agents’ states of knowledge.
- **Miscoordination:** Agents may simultaneously attempt incompatible or duplicate solutions.
- **Segregation of situational awareness:** Differences in access to the environment may render some agents uninformed or impotent.
- **Ultimate authority:** Agents may fail to cooperate, ignore one another, or disagree on the appropriate action.
- **Ground truth to responsibility:** By default, MAS agents may be unaware of the context or purpose of their work, thus completing it less effectively
- **Critical sub-task size:** Benefits to MAS accrue only if the task may be divided into components which are well-defined, isolated, and only moderately complex—simple tasks are better as **tools**.
- **Credit assignment & failure localization:** Some MAS failures are difficult to localize and thus to attribute or fix; relatedly, any form of RL is rendered more difficult.

Remarks on Methodology and Operationalization

Fundamentally benchmarking MAS against SAS poses a design problem: MAS add an additional layer of design beyond SAS, which can strongly influence effectiveness.

Stepping back, the ultimate question: what breadth of tasks can a system complete at low enough cost and high enough reliability to be *useful* toward human goals?

We can think of a few important aspects feeding into these:

- **Speed (wall time):** can running multiple parallel agents get tasks done *faster* than a single agents
 - (Even if less skillfully, parallelizability may dominate.)
- **Token efficiency:** how much compute is expended to solve a given problem?
 - Likely **more** for MAS no matter what.
- **Solution quality:** for problems with graded quality, do we see better or worse solutions?
 - This might include **legibility** (e.g., code).
- **Verification:** do multiple agents permit more robust **intra-task evaluation** (e.g., reduced self-consistency bias)?
- **Distribution of competence:** how do component agents' unique capability profiles contribute?
- **Robustness:** how often do failures occur, how often are they recognized, and how well recovered from?
- **Tool use:** how effectively does MAS utilize tools compared to SAS?
- **Information management:** how do MAS vs SAS cope with large data?

Ideally, experiments would measure as many of these as possible

Appendix C: Elicitory Scenarios

We propose to make an (imperfect) distinction preliminary between two types of systems which we expect to have saliently different development, capability, and deployment profiles:

- **“Weakly multi-agent”**: these systems are essentially single-agent systems with minor provisions made to ensure agents do not ‘step on one another’s toes’. They are best conceived of as single-agents working in the same ‘space’.
- **“Strongly multi-agent”**: these systems derive a significant part of their capabilities from the combination of different component models; they are sufficiently coherent that they are best understood as a ‘unitary agent’.

We are mostly interested in timelines—therefore we focus on the role MAS may play in extending the capabilities frontier. Secondly, we are also interested in possibilities of a rapid expansion of applications - very cheap agentic coding models facilitated by a few middlingly intelligent coordinators might not move the frontier directly but could indirectly accelerate progress via coding automation.

We will take for granted the existence of a sharp distinction, less because it is sharp and true than because it facilitates specificity, between agents and tools: GPT-5 calling DALL-E is an agent calling a tool; Opus 4 dispatching Haiku45 is an agent spawning an agent. Router systems as in GPT5 are borderline, but we will consider them as not MAS in this subsection: they do not have the same strengths and weaknesses, and are likely to develop fairly independently of MAS capabilities.

Channeling our inner AI2027-wannabe, we present two scenarios serving a primarily elicitory and illustrative purpose.

Two Elicitory Scenarios

Weakly multi-agent world: present trends towards small, diverse models continue. Both because of existing infrastructure like git and cursor, and widespread existing verification practices like CI/CD, this makes automated coding much cheaper, though without substantially increasing its quality. Just like empirical studies of CoT abound today, so soon will studies evaluating the best way to generate MAS systems; the present focus on topology and information flow will be outmoded in favor of a hierarchy of agents, with a frontier LLM at the top managing a handful (or many) independent small models. Academics will write papers about how to do optimization over these hierarchical setups, but they will by-and-large be irrelevant — the frontier will still belong to the hulking behemoths, and a gap opens up: labs in possession of frontier models now have to ask how to divvy up resources between pushing forward the frontier (even as the exact location of the frontier is rendered less relevant because O(1T) models are deployed much less) on the one hand, and making their 30B models as close as possible to the 1T models in performance — revenue considerations favour the latter, while even the unreformed transhumanists gradually reconcile their vision with the profit motive (perhaps on the grounds that

X is the great bottleneck stopping rapid automation of AI R&D, where X is today coding, perhaps tomorrow 9's of reliability, etc.). [In this world, what is the comparative standing of Chinese labs — do they have more expertise in model miniaturization? What about applications?] As deployment becomes cheaper, applications open up — would-be robotics AIs might incorporate multiple levels of LLM control and planning — imagine a master LLM communicates to per-limb LLMs which communicate through CLIP-analogous language-action interfaces to specialized motor AIs. While AI becomes much more naturally scalable, there may be little benefit to distilling domain-narrow AIs, at least via distillation — our micro-agents will mostly remain generalists. By contrast, those who have yet the spirit to run 10T params worth of frontier models in a MAS find this acervation to be just as finicky as ever — the agents flail, ignore one another, or are sufficiently misaligned to deny the whole any coherence, with no apparent method to accelerate the whole as RL did the parts. Whether or not the capabilities frontier is accelerated compared to the counterfactual depends largely on how much labs benefit from extremely rapid/cheap generation of 50th percentile code afforded by extremely cheap agentic deployments, and, perhaps, if the increased prospect for deployment revenue staves off the Damoclesian “bursting of the bubble”.

Strongly multi-agent world: Frontier SAS begin to be outperformed by well-designed multi-agent systems. MAS show better wall-clock speed (using faster models for many tasks), superior long-term memory (due to context conservation and effective summarization), and even improved coherence (better context conservation winning out over traditional MAS barriers thereto). This is likely due to an increase in introspective awareness of more advanced models (or of similar models more intensively RLed) — agents know what they don't know, and thus when to phone a friend; likewise, small improvements in long-term coherence of SAS have their impacts magnified by more efficient context usage in MAS; many of the hacky-tacky LLM appurtenances like RAG are replaced by simple queries to nano models finetuned to ingest and summarize oceans of information. As models become introspectively aware, they are increasingly able (perhaps hackily fine-tuned in a mock-MAS setting) to recognize when they're stuck, and spawn an intelligent-but-different peer to take over or dialecticize. There is still plenty of room for explosion of deployment opportunities as smaller models improve — there even sprouts a cottage industry distilling exotic hybrids from frontier LLMs, EBMs, diffusion LLMs, etc. In this world, the bump to capabilities is perhaps appreciable and of wide breadth, but is likely to be a one-off — it is still essentially difficult to optimize the MAS scaffolding directly, so research continues to focus on pushing SAS models along with continued miniaturization efforts. Compared to the counterfactual world, this world sees much more alpha in exploring novel architectures — novel architectures must only outperform the competition in *one* domain (at reasonable cost etc) to earn their keep. Safeteyists continue to wring their hands nervously, some more, others less. Some say multi-agents systems are fundamentally more difficult to align, while others expect that inter-agent collaboration, along with the difficulty of RL-ing a whole MAS at once, should exert an economic pressure towards aligning individual agents. Andrej Karpathy comes up with an insightful analogy of centaur MAS to some computer-science thing, and I nod approvingly when I hear it.

These worlds are not so distinct. The key delta in inputs is how well problems of coherence and introspection are solved by default. The delta in outputs is a change in the makeup of frontier systems, and thus a change in the incentives dictating their preparation. The striking thing to me is that focusing on MAS made me appreciate the likely impact of miniaturization (which will very likely continue apace in our own world). In the first world, it is common wisdom that MAS are too finicky to be worth the effort, while in the second, decades-old MAS practitioners appear on Dwarkesh to praise the achievements of the current paradigm through gritted teeth and unsubtly imply that they were visionaries ahead of their time.

The main vibe-priors that influenced these stories are

1. Frontier model progress becomes increasingly expensive, whether slower or no, so improvements in mundane utility are driven by economization rather than capabilities expansion.
2. Public psychology (via political blowback, widespread disdain for the quest for superintelligence) and market pressures (both the usual, and fears of bubble) push frontier labs to focus resources on rapidly expanding economically-useful deployment.

We consider it is plausible these priors are more load-bearing than the conditioning on the different state of MAS in either.

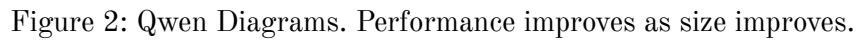
Appendix D. Suggested Forecasting Questions for MAS Capabilities

The focus of this work was to investigate the possibility of a ‘multi-agent’ overhang by explicating the key characteristics of multi-agent systems and investigating whether current models exhibit capabilities enhancements when combined into multi-agent systems. Below, we sketch some forecasting questions which we would have, er, forecasted, had we the time. Most require further specification for resolution, setting of par values, or even resolve based on information likely to be proprietary - thus they are illustrative, and offered, solely as ways to operationalize the questions examined in this paper.

For $X = \text{day of submission} + 3/6/12 \text{ months}$. All other Y and Z left undetermined. Pareto optimality shall be understood to refer to LMSys Elo and per-token-cost.

- By X date, Y percentage of queries to Z model will come from AIs with both higher LMSys Elo and higher per-token cost.
- By X date, Y percentage of queries to Z model will be made by an AI which itself was called by an AI.
- By X date, an off-paradigm frontier model will be Pareto superior to Claude 4 Opus (>1430 LMSys Elo + < 50 \$/M tokens).
- By X date, an off-paradigm frontier model will be Pareto superior to Gemini.
- Conditioning on OAI allowing a single chat-completion to be partially generated by more than one agent, by X $Y\%$ of completions will use models with $>100\times$ difference in active parameters.
- By X , a strongly multi-agent system ($\geq$ one on-paradigm + one off-paradigm) will be top-3 on benchmark Y .
- By X , $Y\%$ of Cursor completions will be generated by a diffusion LLM.
- By X , $Y\%$ of LLM inference compute will be allocated to on-paradigm LLMs with $\leq 10B$ parameters.
- By X , among codebases of $[1k, 10k, 100k] + LOC$, the weekly average number of agents simultaneously active will be Y .
- By X , of the Pareto frontier in LMSys Elo vs \$/M tokens (open-weight only), $Y\%$ will use a router mechanism.

- ## Appendix E. Further Experiments



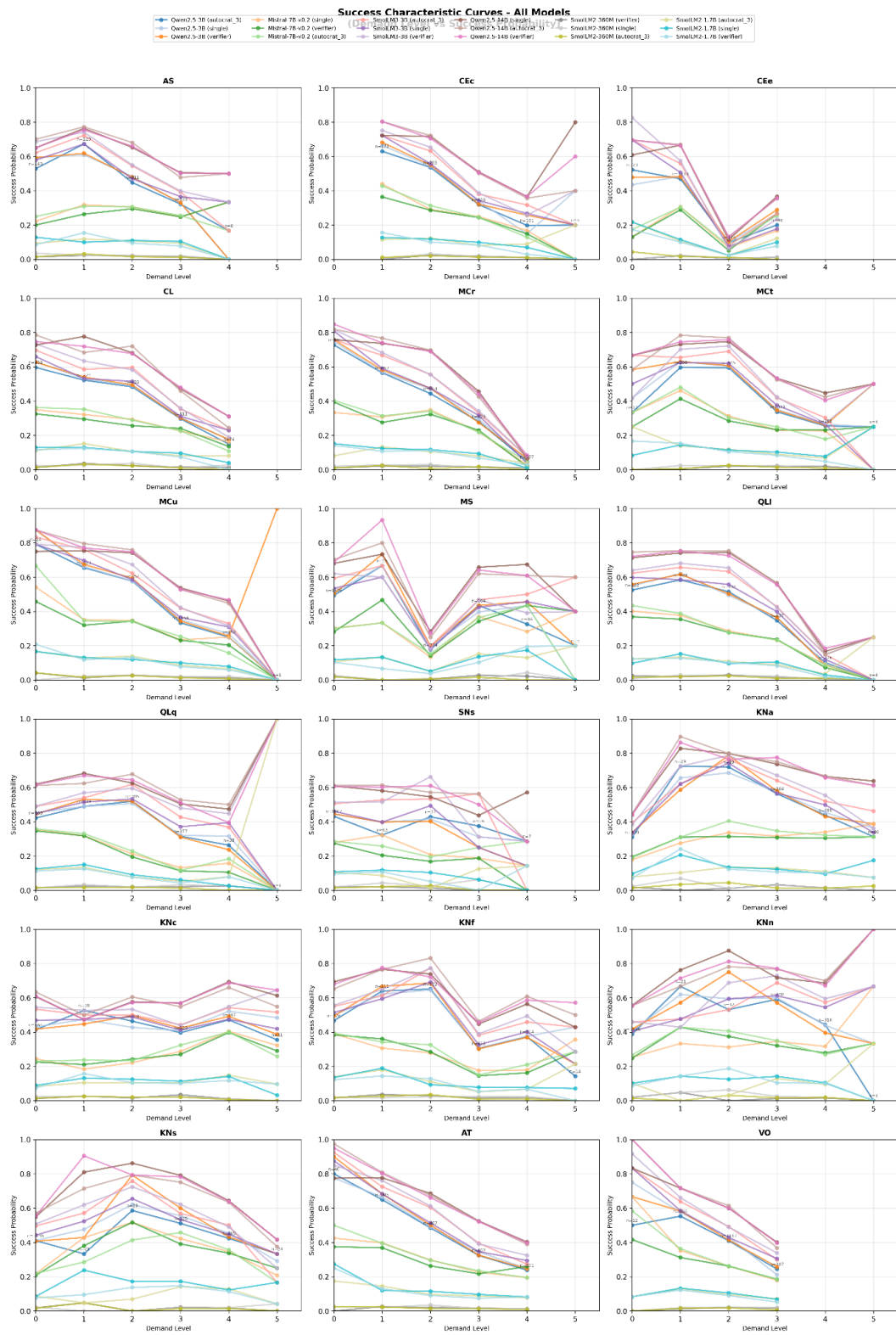


Figure 3. Accuracy vs demand level. How accurate are the models as demand (correlated with difficulty) increase