Fall 2025 AI Benchmarking Participant Survey

This form is intended to gather information about bots in Metaculus's Fall 2025 AI Benchmarking tournament.

- **Survey is required for earning prize:** To earn prize money for your bot you need to fill in this survey.
- **Don't check a box if you barely used something**:
  For example if you only used a model for a week during the competition, please don't list it as a model that you used. We want to find correlations with score, so say that you used/did something only if you used it in a way that noticeably effected your score
- **Results are aggregated anonymously unless otherwise noted**: Aggregated anonymous results will be shared publicly unassociated with your bot. Though to help resolve any potential IP concerns, please assume that anything you share could at some point become public, and answer questions accordingly (notice that not every question is required). If you are a prize winner, you must at minimum fill out enough of the form to constitute a 'description'.
- **Results will be shared with community:** We will run some data analysis on the results and share trends with the bot maker community. By sharing general trends in top bots, our intent is to
  help foster innovation among the wider AI forecasting community and provide useful takeaways to the tournament's funders. By reviewing code we seek to do another level of verification to make sure there is no human-in-the-loop.
- **If you only answer one question:** If you did not win prize money and you only have time to answer one question, please scroll to the bottom and answer the one asking for "what would you share with the world about what you learned". Though note that this means we can't use your bot's data to make better correlations with performance score in the public analysis.

If you feel you can't adequately fill out the form due to IP concerns, or have other questions, please send an email to grace.mclain@metaculus.com and ben@metaculus.com

* Indicates required question

1. What is your bot's name as listed in the Fall 2025 Leaderboard? *

   Please write it verbatim from https://www.metaculus.com/tournament/fall-aib-2025/

2. I confirm that I will answer questions about my bot as it was in the Fall 2025 season *

   Answer yes if so. If you didn't have a bot in Fall 2025, please don't fill in this survey and let us know. We will compare to Fall scores in our analysis.

3. Which LLM model(s) did you use to make your final prediction/answer?

Check all that apply.

- [ ] Finetuned Open Source Model
- [ ] Finetuned Proprietary Model
- [ ] OpenAI o1
- [ ] OpenAI o3
- [ ] o3-mini
- [ ] o4-mini
- [ ] GPT-4o
- [ ] GPT-4.1
- [ ] GPT-4.1-mini
- [ ] Opus 4
- [ ] Sonnet 4
- [ ] Sonnet 3.7
- [ ] Sonnet 3.5
- [ ] DeepSeekR1
- [ ] Gemini-2.5-pro
- [ ] Gemini-2.5-flash
- [ ] GPT 5
- [ ] Claude 4 Sonnet
- [ ] Claude 4.1 Opus
- [ ] GPT 5 Mini
- [ ] GPT 5 Nano
- [ ] Grok 4
- [ ] Kimi K2
- [ ] GPT OSS 120B
- [ ] DeepSeek v3.1
- [ ] Claude 4.5 Sonnet
- [ ] Qwen 3 Max
- [ ] DeepSeek 3.2
- [ ] Grok 4 Fast
- [ ] GPT 5.1
- [ ] Gemini 3 Pro
- [ ] Gemini 3 Flash
- [ ] Grok 4.1 Fast
- [ ] Grok 4.1
- [ ] Other: _____

4. Which LLM model(s) did you use in supporting roles (i.e. not final predictions)?

Check all that apply.

- [ ] Finetuned Open Source Model
- [ ] Finetuned Proprietary Model
- [ ] OpenAI o1
- [ ] OpenAI o3
- [ ] o4-mini
- [ ] GPT-4o
- [ ] GPT-4o-mini
- [ ] GPT-4.1
- [ ] GPT-4.1-mini
- [ ] Opus 4
- [ ] Sonnet 4
- [ ] Sonnet 3.7
- [ ] Sonnet 3.5
- [ ] DeepSeekR1
- [ ] Gemini-2.5-pro
- [ ] Gemini-2.5-flash
- [ ] GPT 5
- [ ] Claude 4 Sonnet
- [ ] Claude 4.1 Opus
- [ ] GPT 5 Mini
- [ ] GPT 5 Nano
- [ ] Grok 4
- [ ] Kimi K2
- [ ] GPT OSS 120B
- [ ] DeepSeek v3.1
- [ ] Claude 4.5 Sonnet
- [ ] Qwen 3 Max
- [ ] DeepSeek 3.2
- [ ] Grok 4 Fast
- [ ] GPT 5.1
- [ ] Gemini 3 Pro
- [ ] Gemini 3 Flash
- [ ] Grok 4.1 Fast
- [ ] Grok 4.1
- [ ] Other: _____

5. How did your bot research questions?

Check all that apply.

- [ ] AskNews DeepNews
- [ ] Other AskNews
- [ ] Exa
- [ ] Tavily
- [ ] Perplexity
- [ ] Anthropic web search
- [ ] XAI web search
- [ ] Gemini web search
- [ ] OpenAI web search
- [ ] A Deep Research Tool
- [ ] Google/Bing/DuckDuckGo Search API (or equivalent like Serp API)
- [ ] Computer Use Model + Web Browser
- [ ] Static web scraping (Only HTML, possibly converted to markdown)
- [ ] Interactive web scraping w/o computer use (Rendering, screenshots, playwright MCP, etc)
- [ ] Social Media scraping
- [ ] Other: _____

6. Did your bot use any of the below forecasting strategies?

Check all that apply.

- [ ] Capping predictions at a max/min
- [ ] Mathmatically calibrating/adjusting predictions based on past forecast data
- [ ] Mathmatically extremizing predictions via code
- [ ] Taking the median/mean/aggregate of multiple forecasts
- [ ] Other mathematical adjustments
- [ ] Used a fine-tuned LLM for research
- [ ] Used a fine-tuned LLM for prediction
- [ ] Check for forecasts on similar Metaculus questions or prediction markets
- [ ] Explicitly calculate/estimate base rates in a rigours way
- [ ] Explicitly run Fermi estimates in a rigours way
- [ ] Explicitly consider consider/categorize future scenarios in a rigorous way
- [ ] Have the LLM do self critiquing or red teaming
- [ ] Generate and then research subquestions
- [ ] Generate and then explicity forecast subquestions
- [ ] Simulating multiple personalities or experts
- [ ] Run generated code
- [ ] Collect and analyze pre-existing datasets
- [ ] Use a library for ochestrating agentic loops (or build your own implementation)
- [ ] Use skills
- [ ] Use tool calling
- [ ] Use subagents
- [ ] Use MCP servers
- [ ] Other: _____

7. What went into the development of your bot?

Check all that apply.

- [ ] LLM finetuning
- [ ] Testing via pastcasting (questions that have already resolved)
- [ ] Testing against community prediction on prediction platforms
- [ ] Testing via generating >50 of your own questions, forecasting them, and resolving them
- [ ] Significant testing via manual review of bot outputs (more than sanity checks)
- [ ] Running multiple bots in parallel to compare results
- [ ] Custom Evals (e.g. testing base rate finding against known answers)
- [ ] Running my bot in MiniBench and using results to inform design decisions
- [ ] Other: _____

8. What best describes you?

Mark only one oval.

- ◯ Hobbyist(s) with professional software experience
- ◯ Hobbyist(s) without professional software experience
- ◯ Commercial Entity
- ◯ Non-Profit Entity
- ◯ Academic Researcher(s)
- ◯ Student(s)
- ◯ Other: _____

9. How many people are on your team?

Please give a number

_____

10. What is your best estimate for how many total active hours (between all team members) have been put into developing your bot?

Mark only one oval.

- 0-8hrs
- 9-15hrs
- 16-40hrs
- 41-80hrs
- 2 full time weeks - 1 full time month
- 1 full time month - 4 full time months
- 4 full time months +
- Other: _____

11. Your best estimate of the number of LLM calls per question?

Mark only one oval.

- 1
- 2-5
- 5-10
- 10-20
- 20-50
- 50-100
- 100+
- Other: _____

12. What is your best estimate of cost per Question?

Mark only one oval.

- $0-$0.09
- $0.1-$0.99
- $1-$2.99
- $3-$4.99
- $5-$9.99
- $10-$19.99
- $20-$49.99
- $50+
- Other: _____

13. When building, have you optimized more for research (external information retrieval) or reasoning (processing information given to the LLM)?

Mark only one oval.

- Strong research lean
- Slight research lean
- About the same for both
- Slight reasoning lean
- Strong reasoning lean
- Other: _____

14. Did you change how your bot predicted questions in Fall compared to Q2?

Mark only one oval.

- Yes
- No
- I didn't participate in Q2
- Other: _____

15. How many iterations of your primary bot did you do that ended up forecasting tournament questions live?

Mark only one oval.

- 0 (I made it and let it loose)
- 1-2
- 3-5
- 6-10
- 11-20
- 21-50
- 51+
- Other: _____

16. If you are a prize winner, provide a way to review your code or a general overview of your bot

Non winners can fill this out as well. Code may be reviewed by someone working on the Metaculus AI tooling. If you choose to provide a general description, if you are a prize winner we may still ask you for a code review, but in this case will have an employee unassociated with our AI initiatives review the code with you. You can invite the GitHub username CodexVeritas to your project if you choose to share code. If possible share a link to a specific git commit snapshot of how your code worked during the competition season.

17. If you included a link to your code or a description, can we share it publicly?

Mark only one oval.

- Yes everything (my bot is open source and the repo is public)
- Only qualitative descriptions of my code as Metaculus sees fit, not the link
- Only my description as written (no code links)
- None
- Other:

18. Can we publicly share your individual survey response in association with your bot outside our aggregated results?

Sharing information about specific bots can help people better understand our benchmark and AI capabilities.

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Only select parts or on special condition (please elaborate below)
- ◯ Other: _____

19. Should we continue running MiniBench?

Check all that apply.

- ☐ Yes, I wouldn't have joined otherwise, it was useful to get started
- ☐ Yes, it is a valuable tool for iterating with my bot
- ☐ Yes, its another chance to earn prize money
- ☐ Yes, other reasons
- ☐ Only if you add more diversity of questions (e.g. LLM generated questions)
- ☐ No, I would not be disappointed if it went away
- ☐ No, other reasons
- ☐ Other: _____

20. In summary, what should other bot makers learn from your experience?

This will be shared verbatim as "Bot maker <random #>" in the report. What do you think worked or didn't work in your bot? What are promising directions you see? How can others build on your work, and what should they not repeat? If you are willing, include links to public code or copied prompts that can give better context.

21. Anything else you want to share?