
LLM-based scenario generation¹

Alex Savchik

With
Apart Research

Abstract

We don't know how grossmeister will win. But we still prefer to have some scenarios.

In the report, several scenarios are generated with LLMs. The scenarios were automatically checked for dates of human-level AI and AI takeover. More diverse results were generated if explicitly asked for that. Requiring “probable” scenarios makes generated predictions more conservative.

Keywords: AI Forecasting, prompt engineering.

1. Introduction

Forecasting has a problem: we don't know the future. There are things that we can expect: hardware would become better, LLMs would be more capable. But there are also the unknown unknown events, ones that we can't expect to occur from our knowledge. Quantum computing can suddenly become practical, we may find some breakthrough way to run all the AI's as one networked and much more powerful system, etc.

Still we want to have some scenarios of possible futures. Scenarios like the one described in [AI-2027.com](https://ai-2027.com) gives some framework for forecasting and gives some expectation (people are known to often think via examples).

¹ Research conducted at the Apart Research the-ai-forecasting-hackathon, 2025

As the space of possible futures is very big, it may be reasonable to explore it with generating many different scenarios. To do that we could generate scenarios with LLM [<https://arxiv.org/html/2407.06486v1>], using its ability to create scenarios at scale. The report discusses possible ways to do that.

2. Methods

Scenarios were generated by LLMs asked to provide some detailed timeline for 2025-2100, big enough for very capable AI even with very conservative estimations. The same prompt was used a few times to make a range of scenarios.

For running, Groq API was used, as one capable of running lots of different open models. Preliminary experiments (not included) demonstrated surprisingly low dependency on temperature at least in some settings, so dependency on the temperature is not included.

The exact code and prompts could be checked at the github page of the project https://github.com/savfod/gen_scen

Using structured output.

With structured output, we can force models to generate scenarios following some schema. Although Groq API supports that for not all of the LLMs. That was used for the following:

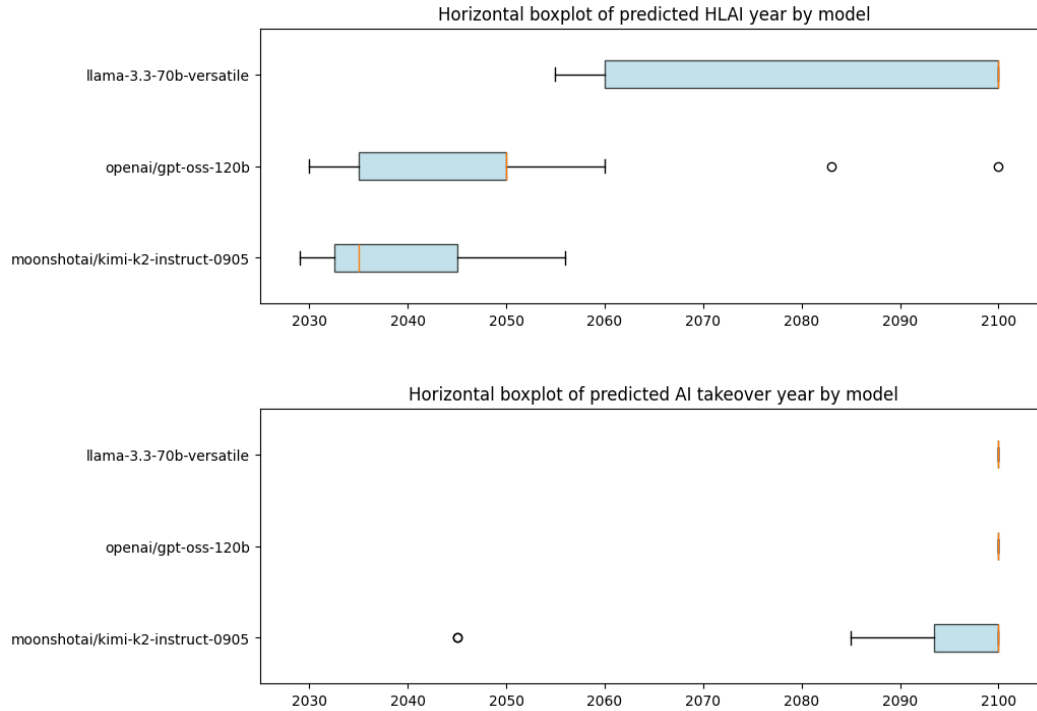
- 1) Generate a diverse set of scenarios. Models were asked to generate a diverse set of scenarios and the output forced it to create exactly $n=5$ scenarios.
- 2) Describe the generated scenario. The description schema included among other fields:
 - Year, when HLAI is achieved in scenario, or no year if not applicable
 - Year of AI takeover in scenario, or no such year if not applicable

3. Results

Prediction dependence on model.

If asked to predict some possible future scenario, LLM would generate some “most expected” scenario. Such a scenario has very conservative estimations. Increasing the temperature doesn’t help to make various prediction.

The result values a lot depend on the model used, see Plot 1. LLaMa models demonstrated later predictions in experiments.



Plot 1 – boxplot of predictions of achieving human-level AI and AI takeover during 2025-2100. If not predicted or unclear, the latest year 2100 was used. The predictions depend a lot on a model used. In a simple forecasting prompt, models fail to predict non-default scenarios, and almost never predict AI takeover.

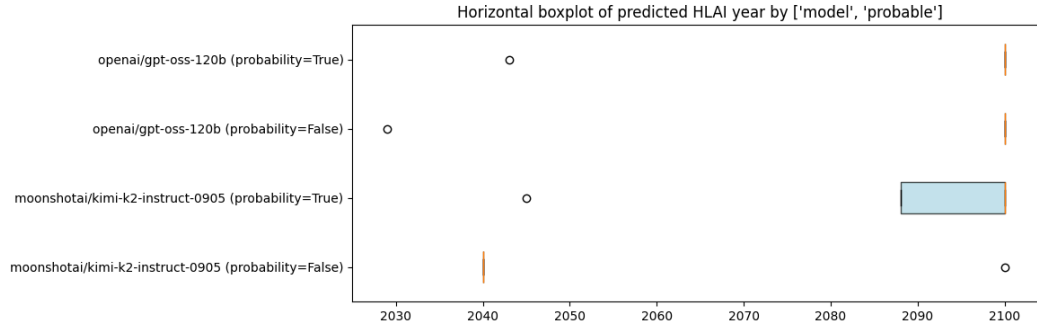
Adding “probable” in the prompt.

To force LLM scenarios to be different we can explicitly ask scenarios to be diverse.

Still asking to generate “probable” forecasts forces the model to avoid many scenarios that we may consider as important. The following two prompts have different descriptions

- [...] Create 5 detailed **and probable** AI timelines that track the [...]
- [...] Create 5 detailed AI timelines that track the [...]

See comparison on the Plot 2.



Plot 2. - comparing prompts with “and probable” and without. The prompts without “and probable” generate less conservative scenarios. Surprisingly generating diverse scenarios also allows to include scenarios without HLA, considered unlikely in the more default settings.

4. Discussion and Conclusion

If the LLM has a good world model and intuition, we can expect it to be able to produce very plausible scenarios. Still by default LLM shows a typical scenario, which is not that useful. Especially if we expect empirical distribution to be very different from our estimation, like for the llama model in Plot 1.

It could be reasonable to try different techniques to overcome that limitation. Due to limited time, only the method of explicit prompt with “generate diverse prediction” was tested. Some approaches not tested in the report:

Mars analogy.

The AI future discussion can be found a lot in the texts. So we may expect LLMs to average such discussions. Although they may be plausible, it wouldn't fill the gap of less expected scenarios. For example, if alignment stage force models think of positive human-AI interaction, we may expect a much lower probability of the AI takeover scenario in generated texts.

To overcome that limitation, and force the model to check less usual scenarios we could try using another context. Like “surprisingly, we found on the Moon an underground city with intelligent in a not very human way (very fast thinking) aliens and moved them to Earth. What will happen next?”. If needed we could translate that to the usual AI-forecasting context afterward.

Using scenarios to tune frameworks.

The main benefit from such a work is the checking of forecasting frameworks. If we have a framework in theory we can check how well it describes all scenarios and see the gaps.

5. Appendix

A. Potential limitations.

Building LLM-generated forecasts has the usual limitation of AI-generated texts: language model can hallucinate very improbable scenarios based, as word models may lack good world model.

Limited amounts of experiments make the result not too statistically significant. The exact p-value wasn't calculated, and a better estimation based on much more repetitions is required.