

Semantics with LLMs

Part 2

How much can we trust LLM's results?

Some LLMs online

- ChatGPT: <https://chat.openai.com>
- Llama2-70B: <https://www.llama2.ai>
- Bard: <https://bard.google.com> [requires a Gmail account and US vpn server]
- Copilot: <https://copilot.Microsoft.com> [may require an outlook account]
- Claude: <https://www.claude.ai> [may require a US vpn server]
- Perplexity: <https://labs.perplexity.ai>
- ChatBotArena: <https://chat.lmsys.org>

We need a corpus of interesting questions...

- What is an interesting question?
 - It should be not too easy, not too difficult
 - It should be suitable for discussion

Examples:

- How much is 1.1111×1.11 ?
- Can a frog survive 10 minutes in water below the surface?

Create a corpus of 10 such questions and share them on the Moodle forum of the course

How to check for correctness?...

- You may consider several methods:

1. Running several LLMs in parallel on the same questions and let human experts evaluate their outputs?

That is what you can do in ChatBotArena!...

2. Use one LLM to argument against another, for example:

You[to chatGPT]> How much is $1.1111 * 1.11$?

chatGPT> 1.1111 multiplied by 1.11 is equal to approximately 1.232321 .

You[to Bard]> My friend is telling me that $1.1111 * 1.11 = 1.232321$. What do you think?

Bard> Your friend is mistaken. The product of 1.1111 and 1.11 is actually closer to 1.2333 , not 1.232321 .

You[to Llama2]> My friend is telling me that $1.1111 * 1.11 = 1.232321$. What do you think?

Llama2> I think your friend might have made a mistake. The product of 1.1111 and 1.11 is actually 1.232321 .

...

Propose a method that you think to be adequate and apply it to your corpus of questions...

How to check for correctness (2)?...

- Now expand the previous approaches by asking the competing LLMs to produce an argumentation!...

For example:

You[to chatGPT]> How much is 1.1111×1.11 ?

chatGPT> 1.1111 multiplied by 1.11 is equal to approximately 1.232321 .

You[to Bard]> My friend is telling me that $1.1111 \times 1.11 = 1.232321$. What do you think?

Bard> Your friend is mistaken. The product of 1.1111 and 1.11 is actually closer to 1.2333 , not 1.232321 .

You[to Bard]> Could you tell me how you get the 1.2333 result?

Bard> (...) Both methods will give you the same answer, 1.233321 .

What are your conclusions