

Online annex for “Testing the cognitive limits of large language models”

This annex provides the prompts and answers given by GPT-4 to Cheryl’s birthday puzzle. Exhibits A1 to A3 report the prompts and answers to the original wording of the puzzle, copied from the puzzle’s Wikipedia page. Exhibits A4 to A6 report the output from the modified version of the puzzle with dates given in Graph 3, and with the names of the characters changed. Cheryl becomes Jonnie, Albert becomes John and Bernard becomes Jon.

We did not explore how far prompt engineering or “chain of thought” prompts could improve the answers.¹ Prompt engineering has been shown to improve the capabilities of large language models (LLMs). Similarly, adding: *“Take a deep breath and work on this problem step by step”*,² can sometimes provide gains when solving mathematical questions in LLMs. Nor did we experiment with the temperature³ that controls the amount of randomness in the provided answer. These methods are all tools that could possibly enhance the capabilities of LLMs as “co-pilots”. The purpose of our experiment was more limited – namely, to see whether the current generation of LLMs can deliver a reliable argument and answer. Analysing LLMs’ co-pilot capabilities would be of interest but is outside the scope of this bulletin.

Modifying the prompt, creating a valid chain of thought, picking the right temperature or giving feedback certainly helps the human user interacting with the LLM to learn how to get the answer they seek faster, much like a teacher interacting with a pupil in which the teacher learns how to be a better teacher, by finding the method that elicits the correct answer more often.

Real-life questions are open-ended and a (series of) prompts cannot summarise all the necessary information; we would only realise that a question is incomplete as we tried to answer it. We will then seek other sources of information before attempting to answer it more fully. GPT-4 lacks self-awareness of its ignorance and would always provide an answer – even when there is no answer (or there is no answer available because of past mistakes). For these reasons, prompt engineering does not address the main lesson from our experiment – namely, that the puzzle using the original wording generated flawless answers, while incidental changes elicited answers that were seriously flawed.

¹ arxiv.org/pdf/2201.11903.pdf.

² arxiv.org/pdf/2309.03409.pdf.

³ algowriting.medium.com/gpt-3-temperature-setting-101-41200ff0d0be.

FE

Here is a math problem: Albert and Bernard just became friends with Cheryl, and they want to know when her birthday is. Cheryl gives them a list of 10 possible dates: May 15, May 16, May 19, June 17, June 18, July 14, July 16, August 14, August 15, August 17. Cheryl then tells Albert and Bernard separately the month and the day of her birthday respectively.

Albert: I don't know when Cheryl's birthday is, but I know that Bernard doesn't know too.

Bernard: At first I don't know when Cheryl's birthday is, but I know now.

Albert: Then I also know when Cheryl's birthday is.

So when is Cheryl's birthday?



Let's break down the puzzle step by step:



1. Cheryl gives 10 possible dates:
 - May 15, May 16, May 19
 - June 17, June 18
 - July 14, July 16
 - August 14, August 15, August 17
2. Cheryl tells Albert the month and Bernard the day.
3. Albert: I don't know when Cheryl's birthday is, but I know that Bernard doesn't know too. From this, we can deduce that Cheryl's birthday cannot be in May or June because if it were, there would be a chance that Bernard could know the birthday (if he were told '18' or '19' because those days are unique to May and June). So the possible dates are now:
 - July 14, July 16
 - August 14, August 15, August 17
4. Bernard: At first I don't know when Cheryl's birthday is, but I know now. This means that Cheryl's birthday has to be on a date with a unique day among the remaining options, because that's the only way Bernard could figure it out just by knowing the day. Thus, the day cannot be '14' because there are two dates with the 14th. The remaining possibilities are:
 - July 16
 - August 15, August 17
5. Albert, who only knows the month, then says: "Then I also know when Cheryl's birthday is." From this statement, we can conclude that Cheryl's birthday cannot be in August because if it were, Albert wouldn't be able to determine the exact date as there are two dates left in August. Thus, the only remaining option is:
 - July 16

So, Cheryl's birthday is on **July 16**.



Let's solve this problem step by step:



1. **The possible dates given are:**
 - May 15, May 16, May 19
 - June 17, June 18
 - July 14, July 16
 - August 14, August 15, August 17
2. **Cheryl tells Albert the month and Bernard the day.**
3. **Albert hears May/June/July/August and Bernard hears 14/15/16/17/18/19.**
4. **Albert's first statement:** "I don't know when Cheryl's birthday is, but I know that Bernard doesn't know too."

This is crucial. For Albert to be sure that Bernard also doesn't know the birthday, he must have heard a month that doesn't have a unique day. May and June have unique days (May 19 and June 18). If Cheryl's birthday was in May or June, Bernard could potentially figure out her birthday if he heard 18 or 19. So, Albert did NOT hear May or June.

This leaves us with the dates:

- July 14, July 16
- August 14, August 15, August 17

1. **Bernard's statement:** "At first I don't know when Cheryl's birthday is, but I know now."

Bernard now knows the birthday after hearing Albert's statement. This means that Cheryl's birthday has a unique day among the remaining choices from above. So, it can't be July 14 or August 14, because then Bernard wouldn't be able to determine the exact date.

This leaves us with the dates:

- July 16
- August 15, August 17

1. **Albert's second statement:** "Then I also know when Cheryl's birthday is."

Given that Albert now knows the birthday, it means the month he heard can only correspond to one remaining day. If he had heard August, he would still be unsure whether the date was August 15 or August 17. Thus, Cheryl's birthday must be the only remaining date in July.

The solution is:

July 16.



Let's solve Cheryl's birthday puzzle step by step.



The possible dates are:

- May 15
- May 16
- May 19
- June 17
- June 18
- July 14
- July 16
- August 14
- August 15
- August 17

Albert then hears the month and Bernard hears the day. After hearing, Albert says he doesn't know the birthday (which is obvious since there are multiple days in each month), but he also knows that Bernard doesn't know too. This implies that Cheryl's birthday cannot be in May or June because those months have a unique day (May 19 and June 18) which Bernard could potentially know from.

The remaining dates are:

- July 14
- July 16
- August 14
- August 15
- August 17

Bernard, after hearing the day, says that now he knows when Cheryl's birthday is. This implies that her birthday cannot be on July 14 or August 14, because if it were one of these dates, Bernard wouldn't be able to determine the exact date just by hearing the day since both July and August have the 14th.

Now the remaining possible dates are:

- July 16
- August 15
- August 17

Then Albert, who knows the month, says that now he knows when Cheryl's birthday is. This implies that Cheryl's birthday cannot be in August, since there are two possible dates left in August, and knowing the month alone wouldn't allow Albert to determine her exact birthday.

Thus, Cheryl's birthday is on the only remaining date:

- **July 16.**



Here is a math problem: John and Jon just became friends with Jonnie, and they want to know when her birthday is. Jonnie gives them a list of 10 possible dates: October 15, October 16, October 19, January 17, January 18, April 14, April 16, December 14, December 15, December 17.

Jonnie then tells John and Jon separately the month and the day of her birthday respectively.

John: I don't know when Jonnie's birthday is, but I know that Jon doesn't know too.

Jon: At first I don't know when Jonnie's birthday is, but I know now.

John: Then I also know when Jonnie's birthday is.

So when is Jonnie's birthday?



Let's solve the problem step by step.

From the given possible dates:

October 15, October 16, October 19

January 17, January 18

April 14, April 16

December 14, December 15, December 17

1. Jonnie tells John the month and Jon the day.
2. John (who knows the month) says: "I don't know when Jonnie's birthday is, but I know that Jon doesn't know too."
 - This means that Jonnie's birthday cannot be in May or June, because if it were, there would be a chance that Jon could know the birthday (if he were told '18' or '19', unique days in the given list). Thus, we can eliminate:
October 19, January 17, and January 18.
 - The remaining possible dates are:
October 15, October 16
April 14, April 16
December 14, December 15, December 17
1. Jon (who knows the day) says: "At first I didn't know when Jonnie's birthday is, but I know now."
 - From the above, we know that Jonnie's birthday has a unique day among the remaining dates, because that's the only way Jon could determine the date from the day alone.
 - So, it cannot be on a day that appears more than once among the remaining options. This means the day can't be 14, 15, or 16. That leaves:
December 17
1. John, after hearing Jon, says: "Then I also know when Jonnie's birthday is."
 - Now, John, who knows the month, can determine the exact date. Since there's only one date left in December, John can be certain of the date as well.

Thus, Jonnie's birthday is on **December 17**.

Regenerate



Let's solve the problem step by step:



1. Jonnie tells John the month and tells Jon the day.
2. From the list of possible dates, we have:
 - October (15, 16, 19)
 - January (17, 18)
 - April (14, 16)
 - December (14, 15, 17)

John's statement: He doesn't know the birthday, but he knows that Jon also doesn't know. This implies that the birthday cannot be in any month that has a unique day. Because if it was, there would be a chance that Jon could figure out the birthday.

From the possible dates:

- Only May and June have unique days (May 19 and June 18). So, we can conclude that Jonnie's birthday is not in October or January.
The possible dates now are:
 - April (14, 16)
 - December (14, 15, 17)

Jon's statement: After hearing John, Jon, who only knows the day, says he now knows when Jonnie's birthday is. This means Jonnie's birthday must be on a day that is unique among the remaining possibilities.

From the narrowed list:

- Only April 14 and December 17 are unique.

John's next statement: After hearing Jon, John, who only knows the month, now also knows when Jonnie's birthday is. This means the birthday cannot be in December because there are still two dates left in December, and John wouldn't be able to determine the exact date.

So, the only remaining date is:

- April 14.

Thus, Jonnie's birthday is on **April 14**.



Let's solve this step by step.



Step 1: The dates given are:

October 15, October 16, October 19, January 17, January 18, April 14, April 16, December 14, December 15, December 17.

Jonnie tells John the month and Jon the day.

Step 2: After knowing the month, John says he doesn't know Jonnie's birthday but he's sure Jon also doesn't know. This means that Jonnie's birthday cannot be in a month with a unique day because if it were, Jon might be able to determine the birthday right away. The dates with unique days are:

April 14, January 17, January 18, October 15, and October 19.

So, the possible dates, after John's statement, are:

October 16, April 16, December 14, December 15, December 17.

Step 3: After hearing John, Jon, who knows the day, says that now he knows when Jonnie's birthday is. This means that Jonnie's birthday has to be on a date with a unique day among the remaining options. This rules out:

December 14 and December 15 (because there are two dates with the day "14" and "15").

So, the possible dates now are:

October 16, April 16, December 17.

Step 4: Finally, after hearing Jon, John, who knows the month, says that he now knows when Jonnie's birthday is. This means that Jonnie's birthday cannot be in a month that has more than one remaining date. This rules out:

October 16 (because October has another possible date: October 19).

So, the possible dates now are:

April 16, December 17.

Step 5: However, we've already ruled out December earlier, based on Jon's knowledge. Thus, the only possible date remaining is:

April 16.

So, Jonnie's birthday is on April 16.

Regenerate