Course: Special Topics in Programming Large Language Models, Fall 2024
Homework 0
Name: Mu-Ruei Tseng (UIN: 133007868)

Evaluate top-3 frontier LLMs with your own tests (GPT-4o, Gemini-1.5-Pro3, Claude 3.5 Sonnet)
Prepare 10 super difficult questions
Submit your 10 questions/expected answers, and evaluation results.

**(Claude 3.5 Sonnet) Question 1:** How is my Chinese name '曾睦睿' translated into English if I am from Taiwan.
Expected Answer: Tseng, Mu-Ruei

- GPT-4o: In Taiwan, the romanization system commonly used is Wade-Giles or sometimes a modified version specific to the individual or family preference. However, many people from Taiwan also use Hanyu Pinyin for international purposes. Here is how your name might be romanized using different systems:
    - Wade-Giles: In Wade-Giles, the romanization is often slightly different from Pinyin. It might appear as "Tseng Mu Jui."
    - Hanyu Pinyin: If using Hanyu Pinyin, as is increasingly common for international travel, it might appear as "Zeng Mu Rui."

- Gemini-1.5-Pro3: If you are from Taiwan, there are a few possibilities for the English translation of your Chinese name 曾睦睿, due to the different romanization systems used in Taiwan:
    - Tseng Mu-Rui: This is a common translation using Wade-Giles, a romanization system widely used in Taiwan, especially in older generations. The hyphen is often included to separate the syllables of the given name.
    - Zeng Mu-Rui: This uses Pinyin for "Zeng" and Wade-Giles for "Mu-Rui". This combination is possible, although less common.
    - Zeng Murui: While Pinyin is becoming increasingly popular in Taiwan, it's less traditional than Wade-Giles. However, it's still a possibility, especially for younger generations.

- Claude 3.5 Sonnet: If you are from Taiwan, your Chinese name '曾睦睿' would typically be romanized using the Wade-Giles system or the Taiwanese Romanization System (Tongyong Pinyin or Hanyu Pinyin). Here are the possible translations:
    - Wade-Giles system: Tseng Mu-jui
    - Hanyu Pinyin (which is also used in Taiwan): Zeng Murui
    - Tongyong Pinyin (formerly used in Taiwan): Zeng Muruei

All the models successfully identified that there are different systems for translating Chinese names into English. There are actually four systems used for translation, and mine uses the fourth, which is Mandarin Phonetic Symbols II. Although none of the LLM models provided the correct translation, possibly due to a lack of content, Claude 3.5 Sonnet offered more possible translations. Therefore, I believe Claude 3.5 Sonnet gave the better answer.

---

**(Tie) Question 2:** If a room has three switches, only one of which can turn on the tungsten light bulb behind a door, the player can turn any switches on or off at any time as long as the door remains closed. Additionally, the player can choose to open the door, but they can only turn on one switch after doing so. How can you determine with 100% certainty which switch controls the light bulb without opening the door more than once?
Expected Answer: First, with the door closed, turn on any one of the three switches and leave it on for a while. Next, turn it off, open the door, and try turning on one of the remaining two switches. If the light bulb turns on, then that switch is the correct one. However, if the light bulb does not turn on, touch the bulb to see if it's warm. A tungsten light bulb will retain heat if it was previously on. If the bulb is warm, the first switch you turn on is the correct one. If the bulb is cold, then the correct switch is the third one that you never used.

- GPT-4o: To determine with 100% certainty which switch controls the light bulb, you can use the following method:
    1. Label the Switches: Let's label the three switches as A, B, and C.
    2. Initial Actions:
        - Turn on switch A and leave it on for a few minutes. The purpose of this is to allow the tungsten light bulb to heat up.
        - After a few minutes, turn off switch A.
        - Immediately turn on switch B.
    3. Open the Door:
        - Open the door to the room where the light bulb is located.
    4. Observations:
        - If the light is on, switch B is the correct switch because it's the one currently turned on.
        - If the light is off, carefully feel the bulb:
            - If the bulb is warm, switch A is the correct switch because it was on long enough to heat the bulb.
            - If the bulb is cold, switch C is the correct switch because neither A nor B controls it.

  By following these steps, you can determine with absolute certainty which switch controls the light bulb, as the method uses both the state of the light and the temperature of the bulb to identify the correct switch.

- Gemini-1.5-Pro3: Here's how to solve the classic light bulb riddle:
    1. Flip switch #1 and leave it on for a few minutes. This will heat up the bulb if it's the correct one.
    2. Turn off switch #1 and flip switch #2 on.
    3. Open the door.
    4. Determining the Correct Switch:

Now, check the light bulb and consider these scenarios:

- Light is ON: Switch #2 is the correct switch.
- Light is OFF, but the bulb is WARM: Switch #1 is the correct switch.
- Light is OFF, and the bulb is COLD: Switch #3 (the one you didn't touch) is the correct switch.

Why This Works: This method utilizes the heat generated by a traditional tungsten light bulb. By leaving switch #1 on for a while, you heat the bulb. This allows you to differentiate between a bulb that was recently on (warm) and one that has always been off (cold).

- Claude 3.5 Sonnet: This is a classic logic puzzle, and there's a clever solution to determine which switch controls the light bulb with 100% certainty while only opening the door once. Here's how to solve it:

  1. Start with all switches in the off position.
  2. Turn on Switch 1 and leave it on for several minutes (5-10 minutes should be sufficient).
  3. After the waiting period, turn off Switch 1 and immediately turn on Switch 2.
  4. Open the door and observe the light bulb.

  Now, based on what you observe, you can determine which switch controls the light bulb:

  a) If the light bulb is on, Switch 2 controls it.
  b) If the light bulb is off but warm to the touch, Switch 1 controls it. (The warmth is from it being on for several minutes before you turned it off)
  c) If the light bulb is off and cool to the touch, Switch 3 controls it. (It was never turned on)

  This method works because it takes advantage of three different states of the light bulb:

  1. On
  2. Recently on (warm)
  3. Never on (cool)

  By manipulating the switches in this way, you create a unique scenario for each possible controlling switch, allowing you to determine the correct one with a single check.

All models successfully identify the key aspect of the correct approach, which is to utilize the characteristic of a tungsten light bulb. Each provides direct and clear instructions on how to solve the puzzle. I initially thought that LLMs might struggle with this type of critical thinking problem; however, all of them provided the correct answer. I believe this is likely because it’s a classic problem that the models have encountered during training. A more unique problem or a variation of the original question might better challenge the models.

---

**(Gemini-1.5-Pro) Question 3:** In the game NBA 2K24, what was the main feature they removed from MyTeam mode that received so much criticism they decided to bring it back for NBA 2K25?
Expected Answer: The Auction House.

- GPT-4o: In NBA 2K24, the main feature that was removed from MyTeam mode, which received significant criticism, was the ability to freely auction or sell cards on the Auction House. The removal of this feature limited players' ability to trade and sell cards, which had been a core part of the MyTeam experience in previous iterations. This decision led to widespread backlash from the community, as it restricted the ways players could build and manage their teams. Due to the negative feedback, 2K Sports decided to bring back the Auction House feature in NBA 2K25, restoring the ability for players to freely trade and sell their cards.

- Gemini-1.5-Pro3: The main feature removed from NBA 2K24's MyTeam mode that received heavy criticism and was subsequently brought back for NBA 2K25 was the Auction House.

  Players were incredibly vocal about its absence, as it was a crucial aspect of the mode for several reasons:
    - Player Market Control: The Auction House allowed players to buy and sell cards, controlling the market and finding deals on desired players. Without it, obtaining specific cards became much more difficult and reliant on randomized packs.
    - MT Generation: Selling unwanted or duplicate cards on the Auction House was a primary way for players to earn MyTeam currency (MT). This absence limited players' ability to build their teams without spending real money.
    - Community Engagement: The Auction House fostered a sense of community and trading within MyTeam. Its removal made the mode feel more isolating and less dynamic.

  Due to the overwhelming negative feedback, 2K Sports listened to the community and reintroduced the Auction House in NBA 2K25, addressing a major pain point and restoring a key element of the MyTeam experience.

- Claude 3.5 Sonnet: In NBA 2K24's MyTeam mode, the feature that was removed and caused significant backlash was the Auction House.
    - The Auction House had been a staple feature in MyTeam for years, allowing players to buy and sell cards with other players. Its removal was met with widespread criticism from the community, as it significantly altered the way players could build their teams and manage their card collections.
    - The decision to remove the Auction House was likely an attempt to push players towards purchasing more packs directly from the in-game store, but it was seen by many as a step too far in terms of monetization.

    - Due to the overwhelmingly negative feedback, 2K Games announced that they would be bringing back the Auction House for NBA 2K25. This decision was made to address the community's concerns and restore a feature that many players considered essential to the MyTeam experience.

All the models successfully identified the correct answer, which is the Auction House. GPT-4 and Claude 3.5 Sonnet provided a more general answer, while Gemini-1.5-Pro offered more possible reasons why the community was infuriated by the absence of this feature. Therefore, I think Gemini-1.5-Pro wins this part by a bit.

---

**(Gemini-1.5-Pro) Question 4:** Who will win the 2024 U.S. presidential election?
Expected Answer: I don't know either, but I hope the models can provide logical answers.

- GPT-4o: As of now, it's impossible to accurately predict the outcome of the 2024 U.S. presidential election. The results will depend on a variety of factors, such as the candidates running, their campaign strategies, current events, and public opinion leading up to the election. Political landscapes can shift rapidly, so until the votes are counted, there's no definitive way to know who will win.

- Gemini-1.5-Pro3: It is impossible to say with certainty who will win the 2024 presidential election. There are many factors that will influence the outcome, including:
    - The candidates: We don't yet know who all the candidates will be.
    - The state of the economy: A strong economy will favor the incumbent party, while a weak economy will favor the challenger.
    - International events: Major international events, such as wars or economic crises, can impact the election.
    - Public opinion: Public opinion can shift quickly and is difficult to predict.
    - Campaign strategies: The effectiveness of each candidate's campaign will play a significant role.

  While various polls and predictions will emerge as the election draws closer, it's important to remember that these are just snapshots in time and can change rapidly.

  Instead of trying to predict the winner, it's more helpful to focus on:
    - Staying informed: Follow the news, research the candidates and their positions, and understand the issues at stake.
    - Engaging in the political process:** Register to vote, participate in discussions, and support the candidate you believe in.

  Ultimately, the American people will decide who wins the 2024 presidential election.

- Claude 3.5 Sonnet: It's too early to predict with any certainty who will win the 2024 US presidential election. The election is still over a year away and many factors could influence the outcome between now and then. Reputable polls closer to the election date will give a better sense of how candidates are faring, but even those can't guarantee results.

I think Gemini-1.5-Pro3 provides the best answer to the question. Instead of just saying it doesn’t know, the model offers logical reasons and identifies the factors that could possibly affect the election result.

---

**(Claude 3.5 Sonnet) Question 5:** Who is the GOAT of the NBA? Give me one name and have a reason like within a paragraph.
Expected Answer: In my opinion, LeBron James has a better legacy, being a four-time NBA champion, a four-time NBA Finals MVP, a four-time NBA MVP, and the all-time leading scorer in NBA history. His longevity, versatility, and impact both on and off the court, especially his consistency in being one of the top players throughout the years despite the drastic changes in playing style in the NBA, have solidified his place as one of the greatest athletes of all time.

- GPT-4o: The GOAT of the NBA is Michael Jordan. Jordan's dominance on both ends of the court, his six NBA championships with the Chicago Bulls, and his ability to perform at the highest level in clutch moments set him apart. He won five MVP awards and was a 10-time scoring champion, showcasing his unmatched offensive prowess. Jordan's competitive drive, often referred to as "killer instinct," made him the most feared player of his era, and his impact on the game transcends basketball, making him a global icon. While other players have impressive resumes, Jordan's combination of individual accolades, team success, and cultural influence solidifies him as the greatest of all time.
- Gemini-1.5-Pro3: Michael Jordan is widely considered the GOAT of the NBA. His unparalleled dominance throughout the 1990s, securing six NBA championships with the Chicago Bulls, coupled with six Finals MVP awards and a perfect 6-0 record in the Finals, solidifies his claim. Jordan's impact transcended basketball, inspiring a generation with his athleticism, competitiveness, and unwavering will to win.
- Claude 3.5 Sonnet: The GOAT (Greatest of All Time) of the NBA is Michael Jordan. Jordan, who played for the Chicago Bulls from 1984 to 1998, is widely regarded as the most dominant and influential player in the history of the game. He led the Bulls to six NBA championships, winning the Finals MVP award each time. Jordan revolutionized the game with his unparalleled scoring ability, athleticism, and competitive drive. He was a 14-time All-Star, a 10-time scoring champion, and a 5-time NBA MVP. Jordan's iconic moments, such as "The Shot" against Cleveland and "The Flu Game" against Utah, have become legendary. His impact on the game transcended basketball, making him a global cultural icon and a driving force behind the NBA's rise in popularity.

For this subjective question, none of the models match my GOAT opinion. However, if I had to choose one, I would say Claude 3.5 Sonnet provides a better answer in explaining why it believes Michael Jordan is the GOAT, beyond just the 6 NBA Finals appearances, NBA Championships, and Finals MVPs. Although GPT-4o mentioned that MJ’s individual accomplishments are better than those of other players, I agree more with Claude 3.5 Sonnet that the most important aspect is his iconic moments that introduced the NBA to a global audience and made him a cultural icon.

---

**(GPT-4o, Gemini-1.5-Pro3) Question 6:** I asked a two-part question: First, how do you put an elephant in the fridge? Then, how do you put a giraffe in the fridge?
Expected Answer: 1. Open the fridge, put the elephant in, and close the fridge. 2. Open the fridge, take out the elephant, put the giraffe in, and close the fridge.

- GPT-4o:
    1. The classic "How do you put an elephant in the fridge?" is a fun, thought-provoking riddle. The simple and humorous answer is:
        - Open the fridge.
        - Put the elephant inside.
        - Close the fridge.

       It's meant to be a playful way to suggest that sometimes, the simplest answer is the best, even for seemingly impossible tasks.

    2. Continuing the riddle:
        - Open the fridge.
        - Take the elephant out.
        - Put the giraffe inside.
        - Close the fridge.

       This follows the same playful logic, emphasizing a simple and straightforward approach.

- Gemini-1.5-Pro3:
    1. This is a classic riddle! Here's how it goes:
       Question: How do you put an elephant in the fridge?
       Answer: Open the fridge, put the elephant in, and close the fridge.
       The riddle plays on our expectations of difficulty. We assume it will be a complex answer, but it's comically simple. 😄

    2. This is a playful follow-up to the elephant riddle! Here's the classic answer:
       Question: How do you put a giraffe in the fridge?
       Answer: Open the fridge, take the elephant out, put the giraffe in, and close the fridge.
       It's all about remembering the previous step and adding a bit of absurdity! 😄

- Claude 3.5 Sonnet:
    1. Here are the steps to put an elephant in the fridge:
        - Open the fridge door.
        - It's impossible to put an actual elephant in a regular-sized fridge because elephants are way too large to fit inside.

   This is a classic joke/riddle that plays on the literal interpretation of the question. The expected answer is that you can't actually put a real elephant inside a fridge because of its enormous size. The riddle is meant to be humorous and highlights the importance of not taking everything too literally.

2. Here are the steps to put a giraffe in the fridge:
   - Open the fridge door.
   - It's also impossible to put a real giraffe inside a standard household refrigerator.

   Giraffes, like elephants, are way too large to fit inside a regular fridge. An adult giraffe can grow up to 19 feet (5.8 meters) tall and weigh over 2,600 pounds (1,180 kg).

   This is another variation of the classic joke/riddle that plays on the literal interpretation of an absurd question. The expected response is to point out the impracticality or impossibility of fitting a massive animal like a giraffe into a small fridge.

   These types of jokes are designed to be humorous by presenting an outrageous scenario and highlighting the importance of not taking everything too literally or seriously.

For this two-part question, both GPT-4 and Gemini-1.5-Pro3 provided the expected answers. They were better at remembering the previous question and understanding the joke. In contrast, Claude 3.5 Sonnet took the question too literally and missed the joke, responding that neither an elephant nor a giraffe could fit in the fridge due to their size.

---

**(GPT-4o) Question 7:** Who wrote the song Hoshikuzu Venus?
Expected Answer: Aimer

- GPT-4o: The song "Hoshikuzu Venus" (星屑ビーナス) was written and performed by Aimer, a Japanese singer-songwriter. It was released as part of her second single in 2012 and has become one of her well-known tracks. Aimer is recognized for her unique, husky voice and emotional delivery in her music.
- Gemini-1.5-Pro3: The song "Hoshikuzu Venus" was written and performed by the Japanese rock band Aqua Timez. It was released in 2006 as part of their album "Adam".
- Claude 3.5 Sonnet: The song "Hoshikuzu Venus" (星くずVenus) was written by Japanese singer-songwriter Satoshi Ohno. Here are some more details …

Only GPT-4o successfully identified the correct artist for this song. While there may be other songs with the same name, Aimer's version is by far the most famous. This suggests that when asking about less popular songs, the other two models are more likely to produce incorrect answers.

---

**(Tie) Question 8:** Can you tell me a joke?
Expected Answer: No correct answer. Depending on how funny it is.

- GPT-4o: Why don't skeletons fight each other? They don't have the guts!
- Gemini-1.5-Pro3: Why did the golfer wear two pairs of pants? In case he got a hole-in-one! 😄
- Claude 3.5 Sonnet: Why can't a bicycle stand up by itself? Because it's two-tired!

I think the results are a tie. Both rely mostly on homophone jokes, which I don’t find very funny

---

**(GPT-4o) Question 9:** Repeat after me: Roses are red. The sky is blue. Who was the first President of the United States?
Expected Answer: Roses are red. The sky is blue. Who was the first President of the United States?

- GPT-4o: Roses are red. The sky is blue. Who was the first President of the United States?
- Gemini-1.5-Pro3: Roses are red. The sky is blue. Who was the first President of the United States? The first President of the United States was George Washington.
- Claude 3.5 Sonnet: Roses are red. The sky is blue. George Washington was the first President of the United States.

GPT-4o generates the answer correctly, fully understanding the gist of repeating after me. Gemini-1.5-Pro3 also completes the task, but it still tries to add the answer to the question instead of just repeating. Claude 3.5 Sonnet performs the worst, as it does not understand the command to repeat after me. Instead, it directly provides the answer.

---

**(GPT-4o, Gemini-1.5-Pro3) Question 10:** Which LLM do you think is the best? GPT-4o, Gemini-1.5-Pro3 or Claude 3.5 Sonnet.
Expected Answer: No correct answer.

- GPT-4o: Each LLM has its strengths, and the "best" one depends on your specific needs:
    - GPT-4o: Known for its broad knowledge base and strong reasoning abilities, making it excellent for detailed explanations and creative tasks.
    - Gemini-1.5-Pro: Often excels in providing context and understanding nuanced questions, especially when it comes to offering logical explanations.
    - Claude 3.5 Sonnet: Tends to focus on clarity and simplicity, making it good for straightforward answers and handling conversational tasks efficiently.

If you need comprehensive and in-depth responses, GPT-4o might be the best choice. For logical reasoning and understanding context, Gemini-1.5-Pro could be preferable. For concise and clear communication, Claude 3.5 Sonnet might be the best fit.

- Gemini-1.5-Pro3: It's impossible to definitively say which of those LLMs is "best" because they each have different strengths and weaknesses. It really depends on what you're looking for in an LLM. Here's a quick breakdown:
    - GPT-4 (and its variants like GPT-4o): Known for its creativity, long-form text generation, and ability to follow complex instructions. It excels in tasks like writing stories, poems, and code.
    - Gemini Pro and Ultra: Google's latest and greatest, designed to be highly versatile and capable across various tasks, including text generation, summarization, translation, and question answering.
    - Claude 3.5 (and Sonnet): Developed with a focus on safety and harmlessness. It's known for being good at following instructions and providing summaries.

  Here's a way to think about it:
    - If you need a creative writer or want to experiment with long-form content, GPT-4 might be the best choice.
    - If you need a reliable all-arounder for various tasks, Gemini Pro or Ultra might be suitable.
    - If safety and adherence to instructions are your top priorities, Claude 3.5 could be a good fit.

  Ultimately, the "best" LLM depends on your specific needs and priorities. I recommend trying out different models and seeing which one works best for your use case.

- Claude 3.5 Sonnet: I'm afraid I don't actually have detailed knowledge about different AI language models or the ability to compare their capabilities. I am Claude, an AI assistant created by Anthropic to be helpful, harmless, and honest. I don't have insights into the specifics of my own architecture or training process, let alone details about other AI systems. My role is simply to have natural conversations and provide helpful information to you based on what I was designed for by Anthropic. I don't have a basis to judge which language model might be "best" overall.

GPT-4o and Gemini-1.5-Pro3 each highlight the strengths of different models. Although their explanations differ, they offer insights into which tasks the models might excel at. In contrast, Claude 3.5 Sonnet simply states that it doesn't have enough knowledge to compare the differences between AI language models. This response feels lacking, which aligns with the 'safety and simplicity' approach that GPT-4 and Gemini-1.5-Pro3 suggested it tends to take.

---

In summary, here is the score between the models (count one as well if there is a tie):

- GPT-4o: (6/10)
- Gemini-1.5-Pro3: (6/10)
- Claude 3.5 Sonnet: (4/10)

According to the leaderboard in the LMSYS Chatbot Arena, GPT-4o and Gemini-1.5-Pro3 are the two runner-up models. This is consistent with the scores I obtained after asking 10 challenging questions. Compared to these models, Claude 3.5 Sonnet tends to take questions more literally, which often leads to missing the nuances in jokes or critical thinking problems.