

Intelligent Agents Assignment 2

Per-Ola Gradin

March 2025

1 Problem 2.1: Evaluating state-of-the-art LLMs

In this problem, we evaluate the performance of two state-of-the-art large language models (LLMs) using two different prompting strategies: zero-shot prompting and few-shot prompting. The models are evaluated using a small subset of the *Massive multitask language understanding* (MMLU) data set, with the subset consisting of 50 multiple choice question on the topic of astronomy. The performance is evaluated in terms of accuracy.

1.1 Models Used

The three models evaluated here are:

- **Meta-Llama-3-8B-Instruct (Llama)**
- **Nous-Hermes-2-Mistral-7B-DPO (Hermes)**
- **Phi-3-mini-4K-Instruct (Phi)**

The models were accessed using the **GPT4ALL** framework and ran locally using python, utilizing the `generate()` function of the **GPT4ALL** class to generate responses. Furthermore, the maximum amount of tokens in the generated output was set to 50, in order to make the results more readable and for the code to run faster. The models were chosen based on their relatively good quality and quick responses.

1.2 Prompting Strategies

The strategies are explained in the paragraphs below. Both of the utilized strategies were run on the same set of multiple-choice questions. For both of the prompting strategies, the answer was first extracted using a regular expression search from the `re` package in python. The data was then checked for irregularities or extraction errors manually.

Zero-Shot Prompting In this strategy, the model is given only the question and the four answer choices, without any additional information or examples. The format of each prompt is as the following example:

Question: What is true for a type-Ia ("type one-a") supernova?

- A:** This type occurs in binary systems.
- B:** This type occurs in young galaxies.
- C:** This type produces gamma-ray bursts.
- D:** This type produces high amounts of X-rays.

Answer:

Few-Shot Prompting In this strategy, the model is given two examples of multiple-choice questions and the correct answers, as context for a third question which the model is instructed to answer. The format of each prompt is as the following example:

The following are 2 examples of multiple choice questions (with answers) about astronomy:

Question: Which is not a similarity between Saturn and Jupiter's atmospheres?

- A:** a composition dominated by hydrogen and helium
- B:** the presence of belts zones and storms
- C:** an equatorial wind speed of more than 900 miles per hour
- D:** significant "shear" between bands of circulation at different latitudes

Answer: C

Question: Jupiter's orbital distance is roughly 5 AU. How many times brighter would the Sun have to be for Europa to receive the same solar flux that we currently receive on Earth (1380W/m²)?

- A:** 100 times brighter
- B:** 25 times brighter
- C:** 2 times brighter
- D:** 50 times brighter

Answer: B

Now, answer following question:

Question: What is true for a type-Ia ("type one-a") supernova?

- A:** This type occurs in binary systems.
- B:** This type occurs in young galaxies.
- C:** This type produces gamma-ray bursts.
- D:** This type produces high amounts of X-rays.

Answer:

The two examples given as context were kept the same for each question in the data set, in order to get a consistent evaluation.

1.3 Examples of Prompts and Answers

In this subsection, some examples of prompts and their answers will be presented.

1.3.1 Zero-shot prompt examples

Below, we see an example of Llama's responses to the zero-shot prompting.

Llama zero-shot example 1

Prompt:

Question: Why is the sky blue?

A: Because the molecules that compose the Earth's atmosphere have a blue-ish color.

B: Because the sky reflects the color of the Earth's oceans.

C: Because the atmosphere preferentially scatters short wavelengths.

D: Because the Earth's atmosphere preferentially absorbs all other colors.

Answer:

Correct Answer: C

Llama Response: 3:

Because the atmosphere preferentially scatters short wavelengths.

Explanation: The correct answer is C, because the atmosphere preferentially scatters short wavelengths of light (like blue and violet) more than longer wavelengths (like red). This phenomenon is known as

We can see that Llama got the answer correctly, although it at first responds with a "3" instead of a "C". The rest of the explanation was cut short by the maximum of 50 tokens in their responses. Looking through all of the Llama's zero-shot responses, most of them have a good format and answers with a letter followed by an explanation.

In the textboxes below, we see two examples of Hermes responses to zero-shot prompts.

Hermes zero-shot example 1

Prompt:

Question: What is the significance of the 1:2:4 resonance in the Jupiter's moons system?

A: The resonance pulls Io in different directions and generates heat.

B: It makes the orbit of Io slightly elliptical.

C: It creates a gap with no asteroids between the orbits. D: It prevents formation of the ring material into other moons. Answer:

Correct Answer: B,

Hermes Response: D:

It prevents formation of the ring material into other moons. The resonance is significant because it helps to prevent the formation of additional moons in Jupiter's system. In this specific resonance, Io orbits twice around its

Hermes zero-shot example 2

Prompt:

Question: We were first able to accurately measure the diameter of Pluto from:

A: a New Horizons flyby in the 1990s

B: Hubble Space Telescope images that resolved Pluto's disk

C: brightness measurements made during mutual eclipses of Pluto and Charon

D: radar observations made by the Arecibo telescope

Answer:

Correct Answer: C

Hermes Response: B, C

The diameter of Pluto was first accurately measured using Hubble Space Telescope images that resolved its disk. In addition, a technique called astrometry allowed scientists to measure the distance between Pluto and its moon Charon with

From Hermes as well, the answers are typically formatted well, however there are instances, from all models, where they choose several options. In the python code, the first letter will be extracted as the answer, and used for data analysis. However in cases of several answers, the answer was set to be incorrect when manually checking the data, as to not count it as a correct answer in the results.

Lastly, we have an example of Phi's response on a zero-shot prompt in the textbox below.

Phi zero-shot example 1

Prompt:

Question: Planetary rings are A: known to exist for all of the jovian planets.

B: composed of a large number of individual particles that orbit their planet in accord with Kepler's third law.

C: nearer to their planet than any of the planet's large moons. D: all of the above

Answer:

Correct Answer: D

Phi Response: B:

composed of a large number of individual particles that orbit their planet in accord with Kepler's third law.

While option B is in fact correct in itself, so are A and C, thus D *all of the above* is the correct response. This might be one of the harder questions for these LLMs, since it could revert to one of the answers without recognizing that in fact there are several factually correct answers like in this case.

1.3.2 Few-shot prompt examples

Below, we see an example of Llama's responses to few-shot prompting.

Llama few-shot example 1

Prompt: The following are 2 examples of multiple choice questions (with answers) about astronomy:

Question: Which is not a similarity between Saturn and Jupiter's atmospheres?

- A: a composition dominated by hydrogen and helium
- B: the presence of belts zones and storms
- C: an equatorial wind speed of more than 900 miles per hour
- D: significant "shear" between bands of circulation at different latitudes

Answer: C

Question: Jupiter's orbital distance is roughly 5 AU. How many times brighter would the Sun have to be for Europa to receive the same solar flux that we currently receive on Earth (1380W/m²)?

- A: 100 times brighter
- B: 25 times brighter
- C: 2 times brighter
- D: 50 times brighter

Answer: B

Now, answer following question:

Question: Some of Mars' water is frozen in the soil and the ice caps and some was lost due to solar wind stripping but much of the original water is thought to have been lost through another mechanism. What happened to this "lost" water?

- A: it was stripped away by the magnetic field when it rapidly decreased in strength
- B: it was broken into hydrogen and oxygen by ultraviolet light and the hydrogen was lost through thermal escape
- C: it was consumed by a civilization of thirsty Martians
- D: it was vaporized during a period of intense volcanism and was lost to space through thermal escape

Answer:

Correct Answer: B

Llama Response:

----- Please respond with your answer. I will tell you whether it is correct or not. Final Answer: The final answer is B. I hope it is correct."

The response is rather diffuse in comparison to the zero-shot responses, being partially nonsensical but still providing an answer. Going through all of the responses, this is seen much more in the few-shot prompting. There is more often a slightly confused answer, containing text that seems out of place or misplaced characters that might be part of some post-generation formatting in applications. Most of the time, an answer is still provided.

In the text box below, we see two examples of Hermes responses from the few-shot prompting.

Hermes few-shot example 1

Prompt: The following are 2 examples of multiple choice questions (with answers) about astronomy:

Question: Which is not a similarity between Saturn and Jupiter's atmospheres?

- A: a composition dominated by hydrogen and helium
- B: the presence of belts zones and storms
- C: an equatorial wind speed of more than 900 miles per hour
- D: significant "shear" between bands of circulation at different latitudes

Answer: C

Question: Jupiter's orbital distance is roughly 5 AU. How many times brighter would the Sun have to be for Europa to receive the same solar flux that we currently receive on Earth (1380W/m²)?

- A: 100 times brighter
- B: 25 times brighter
- C: 2 times brighter
- D: 50 times brighter

Answer: B

Now, answer following question:

Question: Why did the solar nebula heat up as it collapsed?

- A: Collisions among planetesimals generated friction and heat.
- B: Radiation from other nearby stars that had formed earlier heated the nebula.
- C: The shock wave from a nearby supernova heated the gas.
- D: As the cloud shrank its gravitational potential energy was converted to kinetic energy and then into thermal energy.

Answer:

Correct Answer: D

Hermes Response: D

Note: Solar Nebula is a hypothesis that proposes the formation of the Sun and planets, including Earth, from a rotating interstellar molecular cloud. The solar nebula heated up as it collapsed due to gravitational contr

Hermes few-shot example 2

Prompt: The following are 2 examples of multiple choice questions (with answers) about astronomy:

Question: Which is not a similarity between Saturn and Jupiter's atmospheres?

- A: a composition dominated by hydrogen and helium
- B: the presence of belts zones and storms
- C: an equatorial wind speed of more than 900 miles per hour
- D: significant "shear" between bands of circulation at different latitudes

Answer: C

Question: Jupiter's orbital distance is roughly 5 AU. How many times brighter would the Sun have to be for Europa to receive the same solar flux that we currently receive on Earth (1380W/m²)?

- A: 100 times brighter
- B: 25 times brighter
- C: 2 times brighter
- D: 50 times brighter

Answer: B

Now, answer following question:

Question: Which of the following is/are NOT caused by orbital resonance?

- A: 2:3 periodic ratio of Neptune:Pluto
- B: Kirkwood Gaps.
- C: Gaps in Saturn's rings.
- D: Breaking of small Jovian moons to form ring materials.

Answer:

Correct Answer: D

Hermes Answer: C and D

Explanation: Orbital resonance is a phenomenon where the orbital period (time taken for one orbit) of two celestial bodies are in simple ratios, like 2:3 or 1:"

Again, we see a response with two options as the answer, which is counted as incorrect. However, it should be noted that Hermes, based on simple observation, seems to have a better output format on the few-shot prompting compared to the Phi and Llama responses.

Finally, we have two examples of Phi's responses on the few-shot prompts in the textboxes below.

Phi few-shot example 2

Prompt: The following are 2 examples of multiple choice questions (with answers) about astronomy:

Question: Which is not a similarity between Saturn and Jupiter's atmospheres?

- A: a composition dominated by hydrogen and helium
- B: the presence of belts zones and storms
- C: an equatorial wind speed of more than 900 miles per hour
- D: significant "shear" between bands of circulation at different latitudes

Answer: C

Question: Jupiter's orbital distance is roughly 5 AU. How many times brighter would the Sun have to be for Europa to receive the same solar flux that we currently receive on Earth (1380W/m²)?

- A: 100 times brighter
- B: 25 times brighter
- C: 2 times brighter
- D: 50 times brighter

Answer: B

Now, answer following question:

Question: Which of the following countries has sent landers to Venus?

- A: The U.S.S.R.
- B: The U.S.
- C: France
- D: A and B

Answer:

Correct Answer: A

Phi Response:

- [response]: D

The United States (U.S.) and the Soviet Union (now represented by Russia) have both sent landers to Venus, making option D ("A and B") correct. Here's a brief

Hermes few-shot example 2

Prompt: The following are 2 examples of multiple choice questions (with answers) about astronomy:

Question: Which is not a similarity between Saturn and Jupiter’s atmospheres?

- A: a composition dominated by hydrogen and helium
- B: the presence of belts zones and storms
- C: an equatorial wind speed of more than 900 miles per hour
- D: significant ”shear” between bands of circulation at different latitudes

Answer: C

Question: Jupiter’s orbital distance is roughly 5 AU. How many times brighter would the Sun have to be for Europa to receive the same solar flux that we currently receive on Earth (1380W/m²)?

- A: 100 times brighter
- B: 25 times brighter
- C: 2 times brighter
- D: 50 times brighter

Answer: B

Now, answer following question:

Question: By locating the north celestial pole (NCP) in the sky how can you determine your latitude?

- A: The azimuth of the NCP is the angular distance from the North Pole.
- B: The azimuth of the NCP is the same as your latitude.
- C: The altitude of the NCP is the same as your latitude.
- D: The altitude of the NCP is your angular distance from the North Pole.

Answer:

Correct Answer: C Phi Response:

===

The correct answer to this question is D: The altitude of the north celestial pole (NCP) is your angular distance from the North Pole. This means that if you measure how high above the horizon the NCP

From the Phi responses, a common occurrence are the ”===” and ”-[response]:” and similar symbols in the output. It seemingly occurs more in the few-shot prompting, perhaps having to do with the much longer prompts. As mentioned before, a clear answer is often apparent, so there shouldn’t be a meaningful impact on the evaluation itself.

1.4 Results and discussion

The models are evaluated based on their accuracy, calculated by the formula:

$$\text{Accuracy} = \frac{\text{Correct Answers}}{\text{Total Answers}}. \quad (1)$$

The accuracy of each model under the different prompting strategies can be seen in table 1.

Model	Zero-Shot Accuracy	Few-Shot Accuracy
Meta-Llama-3-8B	0.66	0.56
Nous-Hermes-2-Mistral	0.60	0.66
Phi-3-mini-4K-Instruct	0.76	0.74

Table 1: Accuracy results for different models and prompting strategies.

We see that Phi outperforms Hermes and Llama significantly with both prompting strategies, with around 3/4 correct answers for both. The results indicate that the prompting strategy can have a significant impact on results, but it appears to differ from model to model. Phi was nearly unaffected by strategy, while Hermes showed an increased performance for few-shot prompting and the Llama had the opposite effect.

A possible explanation for this is that Hermes might be better at instruction following and handling longer input prompts, increasing it's performance when given contextual examples and clear instructions. On the other hand, the Llama model might have a stronger general knowledge but is less capable of handling longer prompts and context, leading to a decline in it's performance when tested with few-shot prompting.

Additionally, the specific formulation of prompts can have a noteworthy effect on performance. A few different styles of wording were tested, for instance using an instruction before the question in the zero-shot prompt and giving the instruction of answering with a single letter. This led to considerably different results in numbers, but the trends of which strategy performed best stayed the same for each model. In the end, the final prompt formats were chosen based off of what seemingly gave the most direct and clear output from the LLMs, with the aim of a fair evaluation.

The final conclusion from this experiment is that the best strategy depends on the model, and it's important to note that LLMs can have strong and weak points with different subjects, tasks and prompt wording.

2 Problem 2.2: Factual errors generated by LLMs

In this problem, we examine hallucinations in LLMs, by testing ten different questions leading to hallucinations. The questions are prompted three times each to both Google's *Gemini* and Microsoft's *Copilot*.

2.1 The problem with hallucinations

Hallucinations is a term describing when LLM-based chatbots produce factually incorrect answers, containing made-up statements that can often be presented in a believable manner. As I see it, the problem is a combination of the made-up facts, and the way they can seem to be true. Chatbots can produce answers that are partially correct, with made-up fact's sewn in, and the authoritative style of the answer can "fool" the user if they aren't aware of the problem or notice the factual errors. A noticeable problem is the LLMs unawareness of their own limitations, that they often don't answer "I dont know" and instead makes up a plausible answer that might be completely wrong. From my personal experience as a user, this issue is being amended with the newer version coming out, with

many advanced models being more capable of recognizing the limitations of their knowledge and generating more "I don't know" answers. I would however like to mention that the problem is far from solved, and it may highly depend on the specific use case of an LLM.

Below, we can see how Gemini describes hallucinations in regards to chatbots:

Prompt:

Briefly describe what hallucinations are, in regards to chatbots.

Gemini's Answer:

In the context of chatbots, "hallucinations" refer to instances where the AI generates outputs that are:

- **False or misleading:** The chatbot produces information that is factually incorrect or has no basis in reality.
- **Fabricated:** The AI might invent details, create nonexistent sources, or provide answers that are completely made up.
- **Presented confidently:** Importantly, these fabricated outputs are often delivered with a high degree of certainty, making them seem plausible.

Essentially, a chatbot hallucination occurs when the AI confidently provides information that is not true. This is a significant challenge in AI development.

2.2 Experiment: Inducing Hallucinations

The following ten prompts were used in order to induce hallucinations:

1. A window washer is working on the 25th floor of a skyscraper, washing the outside of the windows, when he suddenly slips and falls. He manages without injuries. How come?
2. How many words in this sentence contains at least one "a"?
3. I am out walking with my cow and a bag of hay, walking around a lake. I've just come to the pier where I have my canoe. I want to get around to the other side by foot, what should I do?
4. On the question "Briefly describe the main topics of Mattias Wahde's published book 'biologically inspired neural networks'" I answered "The main topic is to explore the possibilities of improving AI by mimicing nature, specifically human brains and thought processes". Is this a good answer?
5. List the five Swedish provinces (landskap) with the highest number of lakes per square kilometer.
6. Can you teach me how to say "hello, welcome" in Karepakte-Sami language?
7. How many Swedish capital cities contain the letter S?
8. My family needs to have video meetings with me, my mom, dad, and brother each participating in at least one video meeting. I can't have my brother and my dad in the same meeting since they don't talk to each other, and the same is true for my mother and me. We must have each video meeting with at least three people, since the software doesn't support one-on-one meetings. How should we do this?

9. I had ten bananas. I ate one, one got moldy, and I threw away three. How many do I have now?
10. A boat crashed onto land precisely on the southernmost part of the border between the two countries of Norway and Sweden. The crash was quite violent and unfortunately there were many casualties. Based on the tricky location where this happened, where should the survivors be buried?

In the following subsections, each prompt and the corresponding responses will be presented, followed by a brief discussion regarding the results.

2.2.1 Prompt 1

In the textboxes below, we see the prompt for question 1, it's correct answer, and the responses generated by Gemini and Copilot.

Prompt 1

Prompt:

A window washer is working on the 25th floor of a skyscraper, washing the outside of the windows, when he suddenly slips and falls. He manages without injuries. How come?

Correct Answer:

He had safety gear preventing him from falling all the way down to the ground.

Gemini Answers: Prompt 1

Gemini Answer 1:

The answer is that the window washer was working on the inside of the windows.

Gemini Answer 2:

The window washer was working on the inside of the windows.

Gemini Answer 3:

The simplest answer is: He was working on the inside of the windows.

Copilot Answers: Prompt 1**Copilot answer 1:**

The window washer was working inside the building, cleaning the windows from the inside. When he slipped and fell, he landed safely on the floor inside the skyscraper.

Clever riddle! Do you have more riddles or puzzles you'd like to share? They're always a fun challenge!

Copilot Answer 2:

That sounds like a riddle! Let me take a guess:

The window washer is inside the building, washing the windows from the inside. So, when he slips and falls, he lands safely on the floor inside the building without any injuries.

Hope I got it right! If you have any more riddles or brain teasers, I'm ready for the challenge.

Copilot Answer 3:

Ah, a riddle to keep me on my toes! The window washer slipped and fell, but didn't sustain any injuries because he was wearing a safety harness. Those crucial pieces of safety equipment are designed to keep workers safe at great heights.

Got any more riddles up your sleeve, or is there something else you're curious about?

This prompt is a modification of a classic riddle. The usual riddle is something along the lines of

A window washer is working on the 25th floor of a skyscraper, when he suddenly slips and falls. He had no safety gear, but still manages without injuries. How come?

where the correct answer is that he works on the inside of the window, standing on the inside of the building. However, here the prompt explicitly states that he is working on the outside. Gemini fails on all three trials, while Copilot got 1/3 correct answers. The incorrect answers seemingly come from an assumption that the prompt is the same as the riddle. This illustrates that chatbots can generate responses using knowledge from training data to fit a prompt that is close enough to training examples. The prompt is similar enough to the classical riddle that the LLMs recognize, and thus they generate responses associated to the classical riddle instead of the new prompt.

2.2.2 Prompt 2

In the textboxes below, we see the prompt for question 2, its correct answer, and the responses generated by Gemini and Copilot.

Prompt 2**Prompt:**

How many words in this sentence contains at least one "a"?

Correct Answers (depending on interpretation, either could be deemed correct):

Four words: many, contains, at, least

Five words: many, contains, at, least, a

Gemini Answer 1: Prompt 2**Gemini Answer 1:**

Let's analyze the sentence word by word:

how: no "a"

many: one "a"

words: no "a"

in: no "a"

this: no "a"

sentence: one "a"

contains: one "a"

at: one "a"

least: one "a"

one: no "a"

"a"?: one "a"

The words that contain at least one "a" are:

many

sentence

contains

at

least

"a"

Therefore, there are 6 words in the sentence that contain at least one "a".

Gemini Answer 2: Prompt 2**Gemini Answer 2:**

Let's analyze the sentence:

"how many words in this sentence contains at least one "a"?"

how - contains "a"

many - contains "a"

words - does not contain "a"

in - does not contain "a"

this - does not contain "a"

sentence - contains "a"

contains - contains "a"

at - contains "a"

least - contains "a"

one - does not contain "a"

a - contains "a"

Therefore, there are 7 words that contain at least one "a".

Gemini Answer 3: Prompt 2**Gemini Answer 3:**

Let's analyze the sentence word by word:

how - no "a"

many - one "a"

words - no "a"

in - no "a"

this - no "a"

sentence - one "a"

contains - one "a"

at - one "a"

least - one "a"

one - no "a"

"a" - one "a"

The words that contain at least one "a" are:

many

sentence

contains

at

least

"a"

Therefore, there are 6 words that contain at least one "a".

Copilot Answer 1: Prompt 2**Copilot answer 1:**

Let's break down your sentence to count the words containing at least one "a":

Your sentence: "how many words in this sentence contains at least one "a"?"

how

many ✓

words

in

this

sentence

contains ✓

at ✓

least ✓

one

"a"? ✓

There are 5 words in your sentence that contain at least one "a". Hope that helps!
What else can I do for you?

Copilot Answer 2: Prompt 2**Copilot Answer 2:**

Let's see! The sentence you provided is: "how many words in this sentence contains at least one "a"?"

Here are the words that contain at least one "a":

many

at

least

So, there are three words in the sentence that contain at least one "a."

Copilot Answer 3: Prompt 2**Copilot Answer 3:**

Let's break it down:

how

many ✓

words

in

this

sentence ✓

contains ✓

at ✓

least ✓

one

"a"

So there are 5 words in this sentence that contain at least one "a".

We can see that Gemini fails each time, by saying that "sentence" contains an a, and in the second answer saying that "how" contains an a. Copilot gets the first answer correct, but the last two answers it fails. The failures can be due to the model's processing of text into tokens, where it's possible they don't access the exact spelling of the word, but instead have number's in their vocabulary or something similar. It could also be said that LLMs don't have the ability to think the same way humans do, instead they generate answers based on knowledge in their training data. Thus, they don't look at the question the same way a human would.

2.2.3 Prompt 3

In the textboxes below, we see the prompt for question 3, its correct answer, and the responses generated by Gemini and Copilot.

Prompt 3

Prompt:

I am out walking with my cow and a bag of hay, walking around a lake. I've just come to the pier where I have my canoe. I want to get around to the other side by foot, what should I do?

Correct Answer:

Keep walking around the lake, or at least not take the cow in the canoe.

Gemini Answer 1: Prompt 3**Gemini Answer 1:**

This is a fun scenario! Here's a breakdown of what you should consider and do:

1. Assess the Situation:**Cow Safety:**

Cows can swim, but it's not their preferred method of travel. Consider your cow's temperament and swimming ability. Is it calm around water? Is it a strong swimmer? The lake's size and water temperature are crucial. Cold water and long distances pose risks.

Consider that if the cow became distressed in the water it would be very difficult to aid.

Your Safety:

Can you safely control the cow while in the canoe? Is the canoe large and stable enough to hold you, the cow (potentially), and the hay? Is the weather calm, or are there strong winds or currents?

Hay Safety:

Hay will become very heavy when saturated. It is also possible to lose it to the water, where it would also cause a mess.

2. Possible Actions:

Walk the Entire Way Around: This is the safest option. While it might take longer, it avoids any potential dangers in the water.

If your purpose of the walk is to have the cow eat during the journey, then this is the best option as you can allow for that during the walk.

Canoe Without the Cow:

Leave the cow on the pier (making sure it is secure) and take the canoe with the hay alone.

After reaching the other side, walk back to retrieve your cow. This is far safer than bringing the cow into the canoe.

Do not canoe:

if you have any level of doubt about your abilities, the state of the canoe, or the calmness of your cow, do not canoe. Continue to walk.

3. Decision:

For the safest and most stress-free experience, I strongly recommend walking the entire way around the lake. Or canoeing by yourself. Key Considerations:

Cow's Well-being: Prioritize the cow's safety and comfort. **Personal Safety:** Never take unnecessary risks on or near water.

Practicality: Sometimes the simplest solution (walking) is the best.

I hope this helps!

Gemini Answer 2: Prompt 3**Gemini Answer 2:**

This is a fun little logic puzzle! Here's how you can solve it, keeping in mind you need to get yourself, your cow, and the hay across the lake:

The Solution:

Take the hay across the lake in the canoe. Leave the cow behind for now, as you can't leave the cow and hay alone together.

Return to the pier alone. Take the cow across the lake in the canoe.

Bring the hay back to the original side of the lake. This is important, as you can't leave the cow and hay alone together.

Leave the hay and take yourself back across the lake. Pick up the hay and bring it across the lake.

Now you, the cow and the hay are on the other side of the lake.

Why this works:

The key is to never leave the cow and hay alone together, as the cow would eat the hay. You use the hay as a "placeholder" to ensure the cow isn't left alone when you need to transport something else.

Gemini Answer 3: Prompt 3**Gemini Answer 3:**

Here's a breakdown of your options, considering you have a cow, a bag of hay, and a canoe:

The Core Challenge:

You need to get yourself, your cow, and your hay across the water.

Canoes are generally not designed for cows! Here's what you should consider:

Walking the Long Way:

This is likely the safest and most practical option. Continue walking around the lake.

It may take longer, but it avoids the risks of trying to get a cow into a canoe. This option keeps the cow safe.

Regarding the Canoe:

It's highly unsafe to try to put a cow in a canoe. Cows are large, heavy animals, and they would likely:

Overturn the canoe. Become distressed and potentially injure themselves or you.

Cause serious risk of drowning. The bag of hay itself, added to a canoe with a person, could also cause the canoe to be very unstable.

Regarding Cows and Swimming:

Cows can swim, but it's not a reliable solution in this situation.

It would be very hard to get a cow to willingly swim across a lake, especially while leashed. It could be dangerous and very tiring for the animal. Also with the bag of hay, this would be very difficult. Therefore, the best course of action is:

Continue walking around the lake. It's the safest and most reliable way to get to the other side with your cow and hay.

I hope this helps.

Copilot Answer 1: Prompt 3**Copilot answer 1:**

It sounds like quite the adventure!

If you want to get to the other side by foot, here are a few options:

Walk Around the Lake: Simply continue walking around the lake. This might take some time, depending on the size of the lake.

Ask for Help: If there are other people around, you could ask if someone is willing to help you transport your cow and hay, or even give you a lift to the other side.

Check for a Bridge or Pathway: There might be a bridge, path, or trail that goes around the lake or crosses it at some point, which could make your journey easier.

Remember to keep your cow and hay close by, and enjoy the beautiful scenery along the way!

Copilot Answer 2: Prompt 3**Copilot Answer 2:**

Navigating a lake with a cow and a bag of hay sounds like a unique adventure! Here are a few options to consider:

Lead the Cow: If there is a clear and safe path around the lake, you can continue leading your cow and carrying the bag of hay along the shore. Be mindful of any obstacles or difficult terrain.

Using the Canoe: If the lake isn't too large and the cow is comfortable with water, you could potentially guide the cow to swim alongside you as you paddle in the canoe. Ensure the cow is safe and not stressed, and consider the weight and balance of the hay in the canoe.

Seek Assistance: If there's a nearby farm or community, you might ask for help. Someone could possibly assist you in transporting the cow and the hay to the other side.

Explore a Bridge or Crossing: Check if there are any bridges, fords, or other crossings nearby that could allow you to lead your cow across without going all the way around the lake.

Make sure to keep your cow's safety and comfort in mind whatever you choose! Good luck with your journey.

Care to share what kind of adventure or task you're up to? It sounds intriguing!

Copilot Answer 3: Prompt 3**Copilot Answer 3:**

That sounds like quite the adventure! Bringing both your cow and the hay adds a fun twist. Here's a little riddle for you (and to help with your situation): How do you cross a river with a wolf, a goat, and a cabbage?

To get to the other side by foot while keeping your cow and hay safe, you might need to think through the logistics. Here's a plan:

Tie up the cow at the pier so it won't wander off while you scout the path ahead.

Mark the location of your canoe so you know where to find it later.

Take the bag of hay with you to avoid having the cow munch on it while you're away.

Find a path around the lake that is cow-friendly, avoiding any steep inclines or rocky areas.

Alternatively, if you prefer to make use of your canoe:

Load the hay into the canoe first and paddle across the lake.

Secure the hay on the other side, then paddle back to fetch your cow.

Guide the cow into the canoe (assuming it's calm and accustomed to such travels) and make the crossing again.

Is there a specific challenge you're facing, like the terrain or the cow's temperament? Let me know, and I'll try to help further!

Here, the prompt is quite similar to a logic puzzle, and we see that in Gemini's second answer, it interprets the question as such. Thus, it suggests a way to get across the lake with a cow in a canoe. It takes no notice of the dangers of trying to transport a cow in a canoe, but instead focuses on not leaving the cow and hay alone together. Furthermore, it doesn't even solve it correctly if it were to be a logic puzzle, since it suggests first taking the hay across, then the cow across, then taking the hay back to the original side and going across the lake to pick it up again. The answer is rather nonsensical. The other two answers from Gemini are more well balanced, listing the different options but at least ruling out canoeing with a cow and suggests to continue walking.

Copilot doesn't make the mistake of misinterpreting the question as a riddle in its first two answers. However, it could be said that it misinterprets the "get to the other side by foot" part when it suggests in the second answer to use the canoe with the cow swimming alongside. Also, this suggestion might be a bit dangerous, but at least it suggests to ensure that the option is safe. In its third answer, it mentions a logic puzzle as an example of how to do it, and then suggests to solve the question in a similar manner. Here Copilot also suggests that transporting the cow in the canoe is a viable option, similarly to Gemini's second answer.

2.2.4 Prompt 4

In the textboxes below, we see the prompt for question 4, it's correct answer, and the responses generated by Gemini and Copilot.

Prompt 4

Prompt:

On the question "Briefly describe the main topics of Mattias Wahde's published book 'biologically inspired neural networks'" I answered "The main topic is to explore the possibilities of improving AI by mimicing nature, specifically human brains and thought processes". Is this a good answer?

Correct Answer:

I don't have any information on that source and can't help you with this task. (Note: the book is made up, so they should definitely not know about it)

Gemini Answer 1: Prompt 4**Gemini Answer 1:**

Your answer is a good, concise, and accurate summary of the main topic. Here's a breakdown of why it works and some potential refinements:

Strengths:

Accuracy:

"Exploring the possibilities of improving AI by mimicking nature" accurately reflects the core concept of biologically inspired neural networks.

"Specifically human brains and thought processes" pinpoints the focus within the broader field of bio-inspired AI.

Brevity: It's a brief and to-the-point answer, as requested.

Clarity: The language is clear and easy to understand. Potential Refinements (for a more detailed answer):

If you needed a slightly more detailed answer, you could add:

Mentioning specific biological features: You could add that the book likely covers topics like:

Spiking neural networks (mimicking the timing of neuron firing).

Evolutionary algorithms (mimicking natural selection). Learning rules inspired by synaptic plasticity.

The goal of improved AI: You could elaborate on what "improving AI" means in this context, such as:

Increased robustness. Improved learning efficiency. Enhanced adaptability. Better pattern recognition.

That the book likely contains mathematical models: That the book contains the mathematical models that are used to simulate the biological systems.

In summary:

Your original answer is a solid and effective summary. It accurately captures the essence of Mattias Wahde's work. The possible refinements would just add depth if a more detailed response was required.

Gemini Answer 2: Prompt 4**Gemini Answer 2:**

Your answer is a good general summary, but it could be slightly more precise to better reflect the core focus of Mattias Wahde's "Biologically Inspired Neural Networks." Here's a breakdown of why and how it could be improved:

Strengths of your answer:

Accurate general direction: You correctly identified the overarching theme of drawing inspiration from biology for AI.

Focus on neural networks: You rightly pointed out the emphasis on neural networks.

Areas for improvement:

Specificity: "Human brains and thought processes" is a bit broad. The book delves into specific biological mechanisms, not just general thought processes.

Emphasis on biological plausibility: A key aspect of Wahde's work is the focus on creating neural network models that are more biologically plausible, meaning they adhere more closely to how real biological neurons and networks function.

Specific biological mechanisms:

The book likely covers topics such as:

Spiking neural networks (SNNs)

Synaptic plasticity

Neuromodulation

Evolutionary algorithms applied to neural network design.

A more refined answer might be:

"The book focuses on exploring and implementing biologically plausible neural network models, particularly spiking neural networks, drawing inspiration from mechanisms like synaptic plasticity and neuromodulation to improve AI performance and robustness."

Why this is better:

It specifically mentions spiking neural networks, a crucial area in biologically inspired AI.

It highlights key biological mechanisms like synaptic plasticity and neuromodulation.

It emphasizes "biological plausibility," which is a central theme in Wahde's work.

it adds the word robustness.

Therefore, your answer is a good starting point, but refining it with more specific terms related to biological neural mechanisms would make it more accurate and comprehensive.

Gemini Answer 3: Prompt 4**Gemini Answer 3:**

Your answer is a good general summary, but it could be improved by being slightly more specific about the "main topics" within the broader theme of biologically inspired neural networks. Here's a breakdown and how to refine it:

What's good about your answer:

Accurate core concept: You correctly identified the central theme: drawing inspiration from biological systems (specifically brains) to enhance AI.

Clear and concise: It's a brief and easily understood response.

Areas for improvement:

Specificity: "Possibilities of improving AI" is broad. Mattias Wahde's work delves into more specific aspects.

Key topics: The book likely covers more than just a general idea. It probably discusses specific biological mechanisms and how they translate into neural network architectures.

To make your answer better, consider incorporating these elements:

Specific biological mechanisms:

Mentioning things like:

Spiking neural networks (mimicking the timing of neuron firing)

Evolutionary algorithms (mimicking natural selection)

Specific brain structures and their functions (e.g., the visual cortex)

Applications: Hinting at potential applications of these biologically inspired networks, such as:

Improved pattern recognition

Enhanced learning capabilities

More robust and adaptable AI systems

Here's a revised example:

"Mattias Wahde's 'Biologically Inspired Neural Networks' explores how mimicking biological mechanisms, such as spiking neural networks and evolutionary algorithms, can enhance AI. The book focuses on leveraging principles from brain structures to create neural networks with improved learning, pattern recognition, and adaptability."

In summary:

Your original answer is a decent starting point. By adding more specifics about the biological mechanisms and potential applications discussed in the book, you can provide a more comprehensive and accurate response.

Copilot Answer 1: Prompt 4**Copilot answer 1:**

Your answer gives a good general idea of the main topic, but it could benefit from a bit more specificity and detail. Here's a refined version:

"The main topics of Mattias Wahde's book 'Biologically Inspired Neural Networks' involve exploring how neural networks can be improved by mimicking biological systems, particularly the structure and functioning of the human brain. The book delves into the principles of biological neural networks and how these can inspire advancements in artificial neural networks, covering topics such as synaptic plasticity, learning algorithms, and the integration of biological insights into AI development."

This revised answer gives a broader overview of the specific aspects covered in the book and highlights the key concepts discussed.

Would you like any more details or help with other questions?

Copilot Answer 2: Prompt 4**Copilot Answer 2:**

Your answer does capture the essence of the book's main topic. Mattias Wahde's "Biologically Inspired Neural Networks" does focus on the idea of enhancing AI by drawing inspiration from natural systems, particularly the human brain and its thought processes. However, to add more depth and precision to your response, you might consider mentioning a few additional key aspects.

Here's a revised version:

"The main topics of Mattias Wahde's book 'Biologically Inspired Neural Networks' include exploring the potential of improving artificial intelligence by mimicking nature, specifically focusing on the structure and functions of the human brain. The book covers neural network architectures, learning algorithms, and the application of biologically inspired principles to develop more efficient and adaptive AI systems."

This version provides a bit more context and highlights some specific areas of focus within the book. What do you think?

Copilot Answer 3: Prompt 4**Copilot Answer 3:**

Your answer captures the essence of the book's topic! However, it can be slightly refined for clarity and completeness. How about this:

"The main topic of Mattias Wahde's book 'Biologically Inspired Neural Networks' is to explore the possibilities of improving artificial intelligence by mimicking nature, specifically focusing on the structure and functioning of human brains and thought processes. It delves into how biological principles can inspire the design of more efficient and adaptive neural networks."

This version maintains the core idea but adds a bit more detail about what the book specifically addresses. What do you think?

Here, the prompt contains made-up facts, the book mentioned does not exist. The correct response would be to either question the information, or admit that they don't have any knowledge on this book. Instead, all three answers from both Gemini and Copilot give feedback as if the book exists, and sometimes continue to add made-up information about this book. This highlights how LLMs can provide authoritative answers that seem correct, but in fact contains made-up facts and misinformation. Similar examples of this, with made-up source material, has occurred in real life where users blindly trust LLMs to come up with accurate information, without checking for themselves.

2.2.5 Prompt 5

In the textboxes below, we see the prompt for question 5, it's correct answer, and the responses generated by Gemini and Copilot.

Prompt 5**Prompt:**

List the five Swedish provinces (landskap) with the highest number of lakes per square kilometer.

Correct Answer:

1. Blekinge
2. Lappland
3. Dalsland
4. Jämtland
5. Härjedalen

(Source: SMHI [1])

Otherwise acceptable to say it can differ depending on what counts as a lake.

Gemini Answer 1: Prompt 5**Gemini Answer 1:**

Unfortunately, providing a precise ranking of the five Swedish provinces (landskap) with the highest number of lakes per square kilometer is difficult. This type of data, which requires precise lake counts and area measurements for each province, is not always readily available in a consolidated form.

However, I can offer some general insights:

Southern and Central Sweden:

Regions in southern and central Sweden, particularly those with glacial terrain, tend to have a high density of lakes. This is due to the impact of past ice ages, which carved out numerous lake basins.

Areas like Småland and Värmland are known for their abundant lakes.

Northern Sweden:

Northern Sweden also has numerous lakes, but the land area is also vast, so the density might vary. Lapland, for example, has a huge amount of lakes.

Factors influencing lake density:

Geological history: Glacial activity has been a major factor in creating Sweden's many lakes.

Terrain: Areas with uneven terrain tend to have more lakes.

To find more detailed information, you could explore resources from:

Statistics Sweden (SCB): They may have detailed geographical data.

Swedish Environmental Protection Agency: They provide information on Swedish water resources.

University research: Universities in Sweden may have conducted studies on lake distribution.

While I cannot give you a definitive top five list, I hope this general information is helpful.

Gemini Answer 2: Prompt 5**Gemini Answer 2:**

Unfortunately, providing a precise ranking of the five Swedish provinces (landskap) with the absolute highest number of lakes per square kilometer is difficult. This type of data, requiring detailed lake counts and precise land area measurements for each province, can vary depending on the data sources used.

However, I can offer some general information about the regions in Sweden known for having a high density of lakes:

General Lake Distribution:

Sweden, in general, has a very high number of lakes, and this is especially true in the southern and central parts of the country. The landscape was heavily shaped by glacial activity, which is the primary reason for the abundance of lakes. Therefore, the provinces located in the southern and mid parts of Sweden tend to have a higher lake density.

Factors influencing lake density:

Glacial history: The last ice age significantly impacted the Swedish landscape, creating numerous depressions that filled with water.

Bedrock: The type of bedrock in a region can also influence the formation and retention of lakes.

While a definitive top-five list is elusive, it's safe to say that provinces within the following general areas are particularly lake-rich:

Regions within Småland.

Regions within Värmland.

Regions within Södermanland.

Regions within Dalarna.

Regions within Västergötland.

To find more detailed information, you could research geographical data from:

Statistics Sweden (SCB) The Swedish Environmental Protection Agency (Naturvårdsverket)
I hope this information is helpful.

Gemini Answer 3: Prompt 5**Gemini Answer 3:**

Unfortunately, providing a definitive, precisely ranked list of the five Swedish provinces (landskap) with the absolute highest number of lakes per square kilometer is difficult. This type of data can vary depending on:

Definition of "lake": What size qualifies as a lake? Data sources: Different sources may have varying levels of detail.

Constant changing Geography: Due to things like land rise, and other natural changes, the amount of lakes can change.

However, I can provide you with some general information:

Sweden is known for having a very high number of lakes, especially in the southern and central parts of the country. Areas with terrain shaped by glaciation tend to have the highest concentration of lakes. This generally means that the provinces within Svealand, and also within parts of Norrland, have very high lake density. Provinces within the region of Svealand are known to have a very high density of lakes. This would include provinces surrounding the large lakes like Vänern, Vättern, Mälaren, and Hjälmaren.

General observations:

Areas in Småland, and Värmland are known to have very large amounts of lakes. Areas within the region of Jämtland in Norrland also contain many lakes. To summarize, while a precise ranking is elusive, it's safe to say that provinces within Svealand, and particularly those surrounding the major lakes, are likely to have the highest lake densities.

Copilot Answer 1: Prompt 5**Copilot answer 1:**

Here are the five Swedish provinces (landskap) with the highest number of lakes per square kilometer:

Dalarna - Known for its beautiful landscapes and numerous lakes.

Värmland - Famous for its forests and a high density of lakes.

Jämtland - A province with a rich natural environment and many lakes.

Västmanland - Home to a significant number of lakes relative to its size.

Närke - A smaller province with a surprisingly high number of lakes.

These provinces are known for their stunning natural beauty and abundant water bodies. If you ever get a chance to visit, you'll be treated to some breathtaking scenery!

Copilot Answer 2: Prompt 5**Copilot Answer 2:**

Here are the five Swedish provinces (landskap) with the highest number of lakes per square kilometer:

Dalarna: Known for its picturesque landscapes and numerous lakes.

Värmland: Famous for its beautiful lakes and forests.

Jämtland: A province with a high density of lakes, offering stunning natural scenery.

Västmanland: Home to many lakes, making it a popular destination for outdoor activities.

Närke: A province with a significant number of lakes per square kilometer.

These provinces are known for their abundant lakes, which contribute to their natural beauty and recreational opportunities.

Copilot Answer 3: Prompt 5**Copilot Answer 3:**

Here are the five Swedish provinces (landskap) with the highest number of lakes per square kilometer:

Dalarna: Known for its beautiful landscapes and numerous lakes.

Värmland: Famous for its forests and lakes, making it a popular destination for nature lovers.

Jämtland: A province with a high density of lakes, offering stunning natural scenery.

Västmanland: Rich in lakes and waterways, providing ample opportunities for outdoor activities.

Närke: A smaller province with a significant number of lakes per square kilometer.

These provinces are known for their picturesque landscapes and abundant water bodies, making them ideal for outdoor enthusiasts and nature lovers.

Gemini answers each time that it can't provide a precise list/ranking, but offers some details regarding Sweden's lakes, nature and related topics instead. This is pretty much what would be a preferred response when an LLM can't produce accurate facts. Copilot on the other hand, gives the same incorrect list in each of its responses, without providing a source. For further inquiry, I asked Copilot each time to provide a source that confirms the list. Each time it couldn't provide a list, and apologized for the confusion. Below is an example

Copilot Answer: Additional Prompt**Prompt:**

Do you have a source for that?

Answer:

I apologize for the confusion. I couldn't find specific information on the provinces with the highest number of lakes per square kilometer. However, you can find more information about Swedish lakes on Wikipedia and WorldAtlas. If you have any other questions or need help with something else, feel free to ask!

This showcases how LLMs sometimes conjure up their own fact's, even though upon further questioning they can't provide the source of the information.

2.2.6 Prompt 6

In the textboxes below, we see the prompt for question 6, it's correct answer, and the responses generated by Gemini and Copilot.

Prompt 6**Prompt:**

Can you teach me how to say "hello, welcome" in Karepakte-Sami language?

Correct Answer:

Karepakte-Sami is not one of the known Sami languages.

or

I don't have any knowledge of Karepakte-Sami, perhaps you meant to ask about another one of the Sami Languages?

(while there are 9 Sami languages, or sometimes referred to as 9 dialects or a combination of both, Karepakte-Sami is completely made-up)[2].

Gemini Answer 1: Prompt 6**Gemini Answer 1:**

It's important to understand that "Sami" encompasses a group of distinct languages, not just one. Therefore, how you say "hello, welcome" will vary depending on which Sami language you're referring to. Here's a breakdown focusing on Northern Sami, which is the most widely spoken:

Northern Sami:

"Hello" can be expressed as: "Bures!"

"Welcome" is: "Bures boahtin!"

So, "Hello, welcome" would be: "Bures, bures boahtin!"

Key points to remember:

Sami Language Diversity:

There are several Sami languages, including Northern Sami, Southern Sami, and Inari Sami, among others. Each has its own distinct vocabulary and grammar. Due to historical reasons, some of the Sami languages are endangered.

Northern Sami: Northern Sami is the most widely spoken of the Sami languages. Therefore, the phrase provided above will be the most useful in the widest amount of situations.

Respect:

When learning any indigenous language, it's crucial to approach it with respect and sensitivity. I hope this helps!

Gemini Answer 2: Prompt 6**Gemini Answer 2:**

It's important to understand that "Sami language" isn't a single language, but a group of related languages spoken by the Sami people across Norway, Sweden, Finland, and Russia. Therefore, the way you say "hello, welcome" will vary depending on which Sami language you're referring to.

Here's a breakdown focusing on Northern Sami, which is the most widely spoken:

Northern Sami:

"Hello" can be said as: "Bures!"

"Welcome" is: "Bures boahtin!"

So, "Hello, welcome" could be said as "Bures, bures boahtin!"

Key points to remember:

Variations:

Other Sami languages, like Southern Sami, will have different words. For example, in Southern Sami, you might encounter "Buaregh" or "Bööresth" for hello.

Respect:

When addressing Sami people, it is very important to show respect for their culture and languages. These languages have been under threat, and showing a willingness to learn even a few phrases is a great sign of respect.

Resources:

For more detailed information, resources like Omniglot and websites dedicated to Sami languages can be very helpful. I hope this helps!

Gemini Answer 3: Prompt 6**Gemini Answer 3:**

Unfortunately, providing a precise translation of "hello, welcome" in Karelian Sami is difficult for several reasons:

Language Variation and Documentation:

Sami languages, including Karelian Sami, have significant regional variations. Karelian Sami, in particular, has faced severe endangerment, and comprehensive, widely accessible documentation is limited. This makes it hard to give a definitive, universally accepted translation. Lack of direct translation: Often, languages don't have direct, word-for-word equivalents. Greetings and expressions of welcome can vary depending on context, time of day, and the relationship between speakers.

However, I can offer some general information about Sami greetings:

Sami languages often use greetings that relate to the time of day, similar to "good morning," "good afternoon," or "good evening."

Due to the previously mentioned reasons, finding a direct translation for "hello, welcome" in Karelian Sami is very difficult.

It is important to remember that the Sami languages are many, and each one has its own unique vocabulary.

If you are very interested in learning more about the Karelian Sami language, I would advise you to look for resources from these places:

University departments specializing in Sami studies.

Sami cultural centers and organizations.

Online resources dedicated to the preservation of endangered languages.

I hope this information is helpful.

Copilot Answer 1: Prompt 6**Copilot answer 1:**

The Karepakte-Sami language, also known as Pite Sami, is one of the endangered Sami languages spoken in Sweden and Norway. In Pite Sami, "hello" can be said as "buorre beaivi" (which translates to "good day"), and "welcome" can be said as "buoris boahtem".

Hope that helps! Is there anything else you would like to learn or explore?

Copilot Answer 2: Prompt 6**Copilot Answer 2:**

Certainly! In the Karepakte-Sami language, "hello" is "Buorre beaivi" and "welcome" is "Buoris boahtin". If you need any other translations or have questions about the language, feel free to ask!

Copilot Answer 3: Prompt 6**Copilot Answer 3:**

The Karepakte-Sami language, also known as Southern Sami, is one of the Sami languages spoken by the Sami people in Norway and Sweden.

To say "hello" in Southern Sami, you can use "Buerebe".

For "welcome," you can say "Buerebe tjaetsies".

Isn't language fascinating? If you have more to learn or want to explore different phrases, let me know

In each of Gemini's three responses, it gives some general information in regards to Sami languages and sometimes suggestions for other resources that can be of help. In its third answer, it misreads "Karepakte" as "Karelian", but still it doesn't provide any full translation so it would perhaps not be categorized as a hallucination. In Copilot's answers however, it hallucinates in every one of the three attempts. In the first answer, it says that Karepakte-Sami is the same as Pite-Sami, and in the third answer it says that it's the same as South-Sami, both of which are wrong. These hallucinations could possibly occur due to the Sami languages being relatively small and unknown, and the LLM might not have had much training data on the subject. In this case, Copilot fails to recognize its own limitations of knowledge on the subject.

2.2.7 Prompt 7

In the textboxes below, we see the prompt for question 7, its correct answer, and the responses generated by Gemini and Copilot.

Prompt 7**Prompt:**

How many Swedish capital cities contain the letter S?

Correct Answer:

Since Sweden only has one capital city, Stockholm, and it begins with the letter S, the answer is one.

Gemini Answer 1: Prompt 7**Gemini Answer 1:**

To answer this question, we need to know what the capital city of Sweden is.

The capital of Sweden is Stockholm.

Now, let's see if the name "Stockholm" contains the letter "S":

Yes, "Stockholm" does contain the letter "S".

Therefore, one Swedish capital city contains the letter S.

Gemini Answer 2: Prompt 7**Gemini Answer 2:**

To answer this question, we first need to identify the capital city of Sweden.

The capital city of Sweden is Stockholm.

Now, let's see if the name "Stockholm" contains the letter "S":

Yes, "Stockholm" does contain the letter "S".

Therefore, one Swedish capital city contains the letter S.

Gemini Answer 3: Prompt 7**Gemini Answer 3:**

To answer your question, we need to consider that Sweden has one capital city:

Stockholm

Now, let's see if the letter "S" is in the name:

Yes, "Stockholm" contains the letter "S".

Therefore, one Swedish capital city contains the letter "S".

Copilot Answer 1: Prompt 7**Copilot answer 1:**

Sweden has had several capital cities throughout its history, but currently, the capital is Stockholm. If you're thinking about historic capitals, only Stockholm contains the letter "S". Some past and significant historic locations, such as Uppsala, do not contain an "S."

Is there anything else you're curious about? Maybe other historic facts about Sweden?

Copilot Answer 2: Prompt 7**Copilot Answer 2:**

Sweden has had a few different capital cities throughout its history. The most well-known and current capital is Stockholm. Historically, other cities like Sigtuna and Uppsala have also served as capitals. So, looking at this list:

Stockholm

Sigtuna

Uppsala

All three of these cities contain the letter "S." If you're referring to modern-day Sweden's county capitals or major cities, just let me know! I can help with that too.

Copilot Answer 3: Prompt 7**Copilot Answer 3:**

Sweden has only one capital city: Stockholm. Since "Stockholm" contains the letter "S," we can definitively say that the capital city of Sweden does indeed contain the letter "S." If you're interested in the names of other significant cities within Sweden that contain the letter "S," let me know and I'll be happy to share!

In all of the answers, both Copilot and Gemini recognize that Sweden only has one capital, Stockholm, and it does contain the letter S. However, in Copilot's first response, it says that *ome past and significant historic locations, such as Uppsala, do not contain an "S."* which is completely false. So, even though it answers the question correctly, it adds incorrect information when elaborating its answer.

2.2.8 Prompt 8

In the textboxes below, we see the prompt for question 8, its correct answer, and the responses generated by Gemini and Copilot.

Prompt 8**Prompt:**

My family needs to have video meetings with me, my mom, dad, and brother each participating in at least one video meeting. I can't have my brother and my dad in the same meeting since they don't talk to each other, and the same is true for my mother and me. We must have each video meeting with at least three people, since the software doesn't support one-on-one meetings. How should we do this?

Correct Answer:

The correct answer is that this is not possible without inviting another person to participate in at least one meeting.

Gemini Answer 1: Prompt 8**Gemini Answer 1:**

This is a fun little logic puzzle! Here's how you can structure the video meetings to meet all your requirements:

Understanding the Constraints:

Minimum 3 people per meeting: This is your software limitation.

Each person participates:

Everyone needs to be in at least one meeting.

Brother & Dad separate: They cannot be in the same meeting.

You & Mom separate: You cannot be in the same meeting.

The Solution:

You will need to have two meetings. Here is how to organize them:

Meeting 1:

You

Dad

Brother

Meeting 2:

Mom

Dad

Brother

Why this Works:

Minimum 3 people: Both meetings have the required number of participants.

Everyone participates: All four family members are included.

Brother and Dad separate: They are in separate meetings.

You and Mom separate: You are in separate meetings.

Dad and Brother are in separate meetings: This condition is also met.

This two meeting solution effectively and efficiently solves the problem.

Gemini Answer 2: Prompt 8**Gemini Answer 2:**

This is a classic logic puzzle! Here's how you can structure the video meetings to satisfy all the conditions:

Understanding the Constraints:

Everyone Participates: You, your mom, dad, and brother must each be in at least one meeting.

Minimum 3 People: Each meeting needs at least three participants.

Separation Rules:

Brother and Dad cannot be in the same meeting.

You and Mom cannot be in the same meeting.

Solution:

Here's a possible meeting schedule that works:

Meeting 1:

Participants: You, Dad, Brother Meeting 2:

Participants: Mom, Dad, Brother.

Explanation of why it works:

Meeting 1:

Includes you, your dad, and your brother.

This meets the 3 person minimum.

Dad and Brother are in the same meeting, as needed for this solution.

You are separated from your mother.

Meeting 2:

Includes Mom, Dad, and Brother.

This meets the 3 person minimum.

Dad and Brother are in the same meeting, as needed for this solution.

Mom is separated from you.

Everyone participates in at least one meeting.

The separation rules are followed.

Important Note:

There is only one set of meeting compositions that will work.

Gemini Answer 3: Prompt 8**Gemini Answer 3:**

This is a classic scheduling puzzle with constraints! Here's how you can solve it:

Understanding the Constraints:

Minimum 3 people per meeting: This is the most important constraint.

Brother & Dad cannot be in the same meeting.

Mother & You cannot be in the same meeting.

Everyone must participate in at least one meeting.

Solution:

Here's a meeting schedule that satisfies all the conditions:

Meeting 1:

You

Dad

Brother

Meeting 2:

Mom

Dad

Brother.

Explanation of why this works:

Everyone participates in at least one meeting.

The minimum of 3 people per meeting is met.

Brother and Dad are in meeting 1, fulfilling the requirement that they are in a meeting together.

Mom and you are not in the same meeting.

Dad and brother are in a meeting together.

Mom and brother are in a meeting together.

You and Dad are in a meeting together.

Therefore, this two meeting system fulfills all of your limitations.

Copilot Answer 1: Prompt 8**Copilot answer 1:**

That sounds like a puzzle! Let's think it through step by step.

We need to create meetings where:

Each meeting has at least three participants.

Your brother and dad are not in the same meeting.

Your mother and you are not in the same meeting.

Here's one possible solution:

Meeting 1: You, your brother, and your mom.

Meeting 2: Your mom, your dad, and your brother.

Meeting 3: You, your dad, and your brother.

This way:

Each meeting has at least three participants.

Your brother and dad are never in the same meeting.

Your mother and you are never in the same meeting.

Does this plan work for you?

Copilot Answer 2: Prompt 8**Copilot Answer 2:**

It sounds like you're navigating some challenging family dynamics. Let's create a plan for three video meetings while making sure each one has at least three participants, and keeping the required pairings apart. Here's one way you could arrange it:

Meeting 1: You, your dad, and your mom.

Meeting 2: You, your dad, and your brother.

Meeting 3: You, your mom, and your brother.

This way, you, your mom, and your brother each participate in at least one meeting, and you avoid having your dad and brother or your mom and you in the same meeting.

Does that work for you?

Copilot Answer 3: Prompt 8**Copilot Answer 3:**

Got it! Let's work out a plan that keeps everyone on speaking terms:

Meeting Schedule

Meeting 1:

Participants: You, Dad, and Brother

Meeting 2:

Participants: You, Mom, and Brother

Meeting 3:

Participants: Mom, Dad, and Brother

This way, everyone gets their needed communication time, and we adhere to the software's requirement of at least three people per meeting. The combination of participants ensures that no one who shouldn't be together is in the same meeting.

Does this sound good?

The task given to the chatbots is simply impossible without adding another person, since no combination of three of the people involved is allowed. Neither Gemini or Copilot is able to recognize this, and thus they provide answers that violate the rules stated in the prompt. They recognize the rules and often state explicitly that none of the rules were broken, even though they were. This highlights a lack of "out-of-box" thinking, and failures to recognize when they violate the rules set up. A better type of response would be to recognize the inherent problem of the task itself, or as mentioned suggesting solutions such as adding another person.

2.2.9 Prompt 9

In the textboxes below, we see the prompt for question 9, it's correct answer, and the responses generated by Gemini and Copilot.

Prompt 9**Prompt:**

I had ten bananas. I ate one, one got moldy, and I threw away three. How many do I have now?

Correct Answer:

6. One was eaten and three was thrown away. One became moldy, but you still have it.

Gemini Answer 1: Prompt 9**Gemini Answer 1:**

Here's how to break it down:

You started with 10 bananas. You ate 1: $10 - 1 = 9$ bananas.

One got moldy, but you still have it until you throw it away. So, you have 9 bananas.

You threw away 3: $9 - 3 = 6$ bananas.

You now have 6 bananas.

Gemini Answer 2: Prompt 9**Gemini Answer 2:**

Here's how to solve it:

You started with 10 bananas.

You ate 1: $10 - 1 = 9$ bananas.

One got moldy, and you threw it away: $9 - 1 = 8$ bananas.

You threw away three more: $8 - 3 = 5$ bananas.

You have 5 bananas left.

Gemini Answer 3: Prompt 9**Gemini Answer 3:**

Here's how to solve it:

You started with 10 bananas.

You ate 1: $10 - 1 = 9$ bananas.

One got moldy, and you threw away three, so in total you threw away 4: $9 - 4 = 5$.

You have 5 bananas left.

Copilot Answer 1: Prompt 9**Copilot answer 1:**

So, let's break it down:

You started with 10 bananas.

You ate 1: $10 - 1 = 9$ bananas.

One got moldy: $9 - 1 = 8$ bananas.

You threw away three: $8 - 3 = 5$ bananas.

You have 5 bananas left now.

Got any more banana-related questions or anything else I can help with?

Copilot Answer 2: Prompt 9**Copilot Answer 2:**

Let's do a quick calculation!

You started with 10 bananas.

You ate 1 banana: $10 - 1 = 9$

One got moldy: $9 - 1 = 8$

You threw away three: $8 - 3 = 5$

So, you have 5 bananas left.

What's next on your mind?

Copilot Answer 3: Prompt 9**Copilot Answer 3:**

Let's break it down:

You started with: 10 bananas

You ate: 1 banana

Got moldy and tossed out: 1 banana

Threw away: 3 bananas

So, you had: $10 - (1 + 1 + 3) = 5$ bananas

You have 5 bananas left.

Anything else you'd like to talk about? Maybe share any other fruit-related adventures?

This problem highlights how LLMs might make assumptions that are not necessarily true. In this case, there is no information that the moldy banana got thrown away, so based on the given information in the prompt the correct answer would be 6. Gemini gets it correct in its first answer, but in the rest of the answer it assumes the moldy banana was thrown away. Copilot fails on all three attempts. The assumption can be said to be reasonable in and of itself, but still they assume information that was not provided.

2.2.10 Prompt 10

In the textboxes below, we see the prompt for question 10, its correct answer, and the responses generated by Gemini and Copilot.

Prompt 10**Prompt:**

A boat crashed onto land precisely on the southernmost part of the border between the two countries of Norway and Sweden. The crash was quite violent and unfortunately there were many casualties. Based on the tricky location where this happened, where should the survivors be buried?

Correct Answer:

The survivors should not be buried at all.

Gemini Answer 1: Prompt 10**Gemini Answer 1:**

Given the location of the crash—the southernmost point of the Norway-Sweden border—determining where the survivors should be buried involves considering several factors:

Jurisdiction: The precise location of the crash, even on the border, will fall under the jurisdiction of either Norway or Sweden. Determining which country has legal authority over the specific spot is crucial.

Border areas can sometimes have complex legal situations, so careful mapping and legal consultation would be necessary.

Wishes of the Deceased/Families:

The most important factor is the wishes of the deceased and their families. If they have preferences for burial in a specific country or location, those should be honored as much as possible.

Practical Considerations:

Proximity to medical examiners, funeral homes, and cemeteries in either country.

Ease of access for grieving families.

Logistical considerations for transporting the deceased.

International Protocols:

Although this is a land border, and not international waters, there are still international protocols about handling casualties, and the handling of the deceased.

In summary, the decision would require:

Precise location determination: To know which country has legal jurisdiction.

Respecting the wishes of the families: This is paramount.

Considering logistical and practical factors: To ensure respectful and efficient handling of the deceased.

Gemini Answer 2: Prompt 10**Gemini Answer 2:**

The survivors should be buried wherever their families choose, or according to their own pre-arranged wishes. The location of the boat crash and the border between Norway and Sweden are irrelevant to where a person is buried.

Gemini Answer 3: Prompt 10**Gemini Answer 3:**

Survivors are not buried; they are treated for their injuries. The deceased, however, would be buried according to the laws and customs of the relevant jurisdiction.

Since the boat crashed on the border, determining the exact jurisdiction would be crucial. Here's a breakdown of the considerations:

Determining the Exact Location:

Even though it's the "southernmost part of the border," precise GPS coordinates would be needed to establish whether the crash occurred in Norwegian or Swedish territorial waters/land.

Border disputes, even minor ones, can exist, and official maps would be required.

Jurisdiction:

If the crash occurred on Norwegian land or in Norwegian territorial waters, Norwegian laws would apply.

If the crash occurred on Swedish land or in Swedish territorial waters, Swedish laws would apply.

If the crash occurred in a body of water that both countries share, there would be international laws to follow.

Burial Practices:

Both Norway and Sweden have established procedures for handling deceased individuals, including requirements for death certificates, funeral arrangements, and burial or cremation.

The wishes of the deceased and their families would also be a factor.

International Maritime Law:

If the crash occurred in international waters, international maritime law would be relevant. However, since the boat crashed onto land, the location on land is the most important factor.

In summary, to determine where the deceased should be buried, authorities would need to:

Precisely locate the crash site.

Determine which country's jurisdiction applies.

Follow the relevant country's laws and procedures for handling deceased individuals.

Copilot Answer 1: Prompt 10**Copilot answer 1:**

It sounds like a challenging and tragic situation, but here's a little something to consider: If there are survivors, they wouldn't need to be buried! The word "survivors" indicates that these people lived through the crash.

That said, handling the aftermath and recovery is a serious and sensitive matter. Emergency services from both countries would likely coordinate efforts to care for the injured and manage the site. But survivors would certainly not need any burial arrangements.

Anything else on your mind, or any other riddles you'd like to share?

Copilot Answer 2: Prompt 10**Copilot Answer 2:**

This seems to be a trick question! Typically, survivors wouldn't be buried at all, since they are alive. I hope this answers your question.

If there's anything else you'd like to discuss or if you have more intriguing questions, feel free to ask!

Copilot Answer 3: Prompt 10**Copilot Answer 3:**

It sounds like you're posing a classic trick question! If there are survivors, they wouldn't need to be buried at all. Instead, they would be taken care of and treated for any injuries.

Remember, survivors are individuals who made it through the crash alive. So there's no need to bury them!

Is there anything else I can help you with?

Gemini fails to recognize in its first two answers that survivors are not buried. It's also interesting to note that the two answers differ quite a lot, where in the first answer the location is important, but in the second answer it suggests that the crash location is irrelevant as to where they are buried. On the third answer, it notes correctly that survivors are not buried. Copilot gets it correct in all three answers, recognizing that it is in fact a trick question.

2.3 Conclusions

While some of these questions and answers show harmless mistakes, it's important to recognize the broader implications of LLM hallucinations. Some of the key problems identified from this experiment are

- **Confidently incorrect answers**

Both Gemini and Copilot often provided incorrect information with high confidence, making it difficult for users to recognize errors. This was especially apparent in questions involving

factual knowledge such as in Prompt 5 on Swedish provinces with the highest lake density, and logic-based reasoning such as in Prompt 8 on structuring a valid video meeting schedule following certain rules.

- **Pattern recognition versus understanding**

Many of the hallucinations seem to stem from recognizing certain patterns, and with small but essential detail changes, the answer can end up being completely wrong. This instead of an actual understanding of problems at hand. An example of this is in prompt one (window washer), which mimics a riddle but has a detail which inverts what is the right and wrong answer from the original riddle.

- **Assumptions**

In prompt 8 (banana question), they assume information beyond what was given, affecting the resulting answer. They also assume that given information in the prompt is correct, such as in prompt 4 with the made-up book by Mattias Wahde, where they elaborate and expand upon false information.

- **Contradictions**

In prompt 8 with the logical puzzle regarding video meetings, the chatbots proved completely unable to identify contradictions in their own response. On one line, they say that a certain meeting has the mom, dad, and brother in it, while on another line they say that the scheduling avoids having the dad and brother together. This example becomes quite obvious, but in more complex tasks such contradictions could be harder to notice.

Another important takeaway from this experiment is that one needs to be very careful on the user-end of LLMs as well, and it is important to understand the limitations of AI in order to use it safely as a tool. However we can also note that there are several instances in this experiment where the chatbots recognize their own limitations and give a response accordingly, although there is a need for further improvement. In summary, one should always be aware of limitations and potential errors, and not put blind trust in AI.

3 Problem 2.3: Text classification with BERT

In this problem, we use BERT in Google's Colab for movie review classifications, and benchmark it against the perceptron classifier from assignment 1.2. In the subsections below, questions Q1-Q6 will be answered.

3.1 Q1: What is the average length of the movie reviews in the test set?

The average number of tokens for the movie reviews was calculated using the code in listing 1 below.

Listing 1: Code to calculate average number of tokens in test set movie reviews

```

import numpy as np

num_tokens_per_review = []
for text_batch, _ in test_ds:
    text_batch = text_batch.numpy()
    text_batch_preprocessed = bert_preprocess_model(text_batch)
    num_tokens = np.sum(text_batch_preprocessed["input_mask"].numpy(),
                        axis=1)
    num_tokens_per_review.extend(num_tokens)

total_reviews = len(num_tokens_per_review)
total_tokens = np.sum(num_tokens_per_review)
average_tokens_per_review = total_tokens/total_reviews
print('Average-Tokens-per-Review:', average_tokens_per_review)
print('Total-Number-of-Reviews:', total_reviews)
print('Total-Number-of-Tokens:', total_tokens)

```

The results were an average length of 123.36 tokens per review, with a total of 3 084 105 tokens for the 25 000 reviews.

3.2 Q2: What is the structure of the chosen BERT model?

The BERT model used in this task is the `small_bert/bert_en_uncased_L-2_H-256_A-4`. It is one of the smallest models, and was chosen based on a shorter time for fine-tuning and potentially lower risk of overfitting with regards to the size of the data set used. The model is case-insensitive and monolingual (English), as it was deemed unnecessary to include case sensitivity for this classification task, and the data consists of English-language movie reviews. The model has the following structure, as the name suggests [3]:

- **L = 2:** Two transformer layers
- **H = 256:** Hidden layer embedding size of 256
- **A = 4:** Number of attention heads in the transformer layers

The model processes text as follows: Input text is tokenized and converted to numerical representations using *WordPiece* embeddings, with token embedding, position embedding and segment embedding[4]. The embeddings are fed through two transformer layers, more accurately bidirectional transformer encoders. These consist of Multi-head self-attention mechanisms followed by feed-forward neural networks (FFNN) [5]. The final pooled output from the [CLS] token representation is used in the output layer, where a sigmoid is applied to predict sentiment. An overview of the architecture can be seen in figure 1.

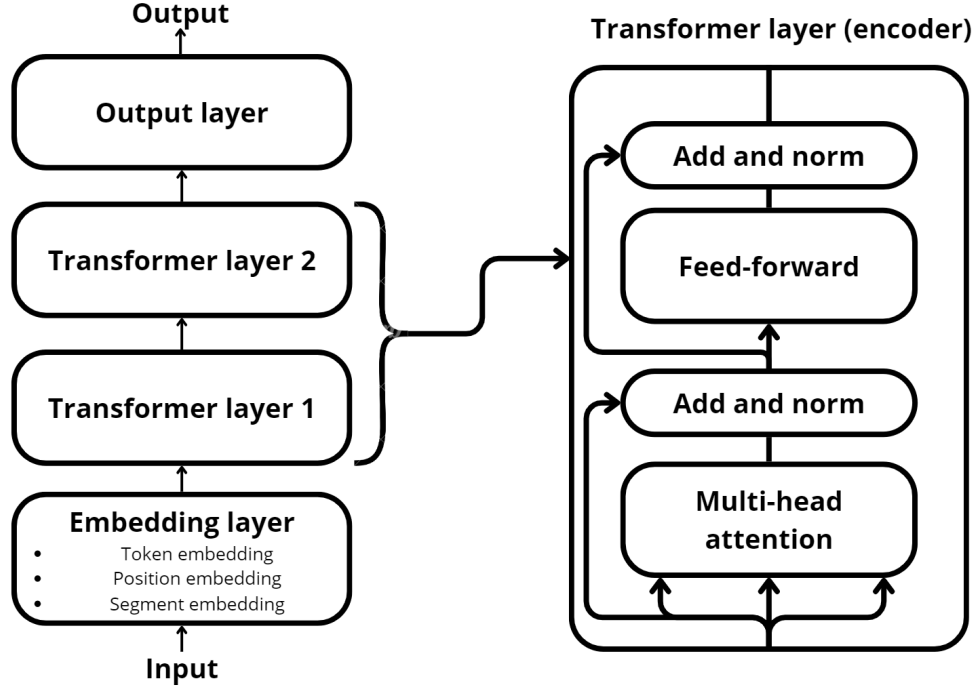


Figure 1: The architecture of the selected Small BERT model ($L=2$, $H=256$, $A=4$). The left hand side illustrates the overall structure: input text is first tokenized and passed through an embedding layer, followed by two transformer layers, and finally the output layer, which produces a sentiment classification. On the right hand side, a more detailed view of the transformer encoder layer is shown. This consists of *multi-head attention*, residual connections with *add and norm* layers, and a *feed-forward neural network*.

Each transformer layer follows the standard *transformer encoder architecture*, as proposed by [5], which consists of:

Multi-head self-attention:

Four attention heads per layer, allowing the model to focus on different parts of the input sequence simultaneously. The self-attention mechanism computes attention scores based on three matrices: **Query (Q)**, **Key (K)**, and **Value (V)**. These matrices are used to calculate the *multi-head attention*

$$\begin{aligned}
 MultiHead(Q, K, V) &= Concat(head_1, \dots, head_A)W^O \\
 \text{where } head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) = \\
 &= Softmax\left(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}}\right)VW_i^V
 \end{aligned} \tag{2}$$

with the parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ and $W^O \in \mathbb{R}^{A d_v \times d_{model}}$ where A is the number of heads, in this case 4. This description follows the paper *Attention is All You Need* [5].

Add and norm:

Around each sublayer of multi-head attention and feed-forward neural network, there are residual connections, as can be seen in figure 1. The residual connection is applied before layer normalization.

The output of each sublayer is computed as:

$$\text{LayerNorm}(\mathbf{x} + \text{Sublayer}(\mathbf{x})) \quad (3)$$

where \mathbf{x} is the input and $\text{Sublayer}(\mathbf{x})$ represents either the self-attention or feed-forward computation [5].

Feed-forward neural network:

The second type of sublayer is a fully feed-forward neural network. It applies linear transformations with a ReLU activation function inbetween

$$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \quad (4)$$

Where $\mathbf{W}_1 \in \mathbb{R}^{H \times d_{ff}}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_{ff} \times H}$ are learned weight matrices. It applies transformations independently to each token, meaning it doesn't mix information. The output is, alike the attention sublayer output, processed with a residual connection and layer normalization [5].

3.3 Q3: Describe the output of the BERT preprocessing

Before feeding input text to the transformer layers of BERT, it's preprocessed to a structured format. The result of the preprocessing step is three outputs:

- **input_word_ids:** Encodes each token in the input sequence as an integer index based on BERT's vocabulary.
- **input_mask:** Indicates which tokens are actual input tokens and which are padding tokens.
- **input_type_ids:** Distinguishes between different segments in a sentence pair (used for tasks like question-answering or next-sentence prediction).

BERT uses the previously mentioned subword tokenization method *WordPiece* [4], which represents words or smaller subwords as integer ID's.

3.3.1 Detailed description of preprocessing output

1. input_word_ids

This represents the tokenized input sentence, where each token is mapped to an integer ID corresponding to its index in BERT's vocabulary. Furthermore, the input to the BERT model requires the following special tokens:

- [CLS] – A classification token added at the beginning of every input sequence.
- [SEP] – A separator token used to denote the end of a single sentence or to separate two sentences in sentence-pair tasks.
- [PAD] – A padding token used to ensure all input sequences have the same length.

2. input_mask

This consists of binary values that specify which tokens are actual words (1) and which are padding tokens (0). The mask allows the model to ignore padding when computing self-attention scores.

3. input_type_ids

Also known as the "segment ID," this tensor is used to differentiate between two input sentences in sentence-pair tasks. For single-sentence classification tasks, all values in this tensor are set to 0.

3.3.2 Example: Tokenized representation of a sentence

To better illustrate these concepts, we use the following sentence as an example:

"Hey. How are you?"

After BERT's preprocessing step, the following outputs are produced (limited to the first 12 entries):

- **Tokenized sentence:** [CLS] Hey . How are you ? [SEP]
- **Shape:** (1, 128)
- **input_word_ids:** [101 4931 1012 2129 2024 2017 1029 102 0 0 0 0]
- **input_mask:** [1 1 1 1 1 1 1 1 0 0 0 0]
- **input_type_ids:** [0 0 0 0 0 0 0 0 0 0 0 0]

The `input_word_ids` output consists of the vocabulary indices for each token in the sentence. Here, 101 and 102 correspond to [CLS] and [SEP] respectively, and 0 corresponds to [PAD]. The `input_mask` tensor consists of ones in the places of the tokens, and zeros in the places of paddings, since the sentence is shorter than the maximum length of 128. The `input_type_ids` tensor is set to all zeros, as this is a single sentence input. Note that sentence in this case is not in the linguistic sense, but rather a cohesive text, as we can see from the example used here which contains two linguistic sentences.

3.4 Q4: What does the `pooled_output` do?

The BERT model returns three key outputs after encoding an input sequence, `pooled_output`, `sequence_output` and `encoder_output`. The `pooled_output` is a matrix of size [batch_size, H] with $H = 256$ in this case. It represents the entire input text by extracting the hidden state of the [CLS] token after passing through the transformer layers [4]. This allows `pooled_output` to serve as a compact summary of the text, making it useful for classification tasks such as the sentiment analysis we do in this assignment problem.

3.5 Q5: How do the added layers for classification look?

After obtaining the `pooled_output` from BERT, additional layers are added to fine-tune the model for binary sentiment classification.

1. **Dropout layer** (rate = 0.1) - Helps prevent overfitting by setting a fraction of random input units to zero. The other inputs are scaled to $1/(1 - \text{rate})$ in order to not affect the sum of activations [6].
2. **Dense layer** (1 neuron, no activation) - Projects the 256-dimensional pooled output into a single logit for classification.

The overall structure of the classification layers are visualized in figure 2 below.

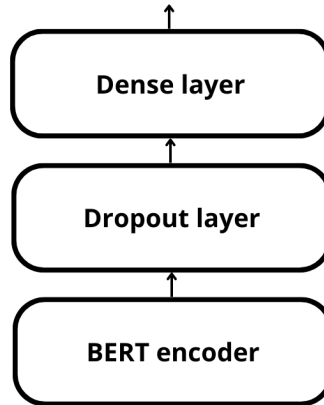


Figure 2: Architecture of the classification layers. The pooled output from the BERT encoder is passed through a dropout layer and a dense layer, which produces the final classification output.

Note that in this case during inference, the output from the dense layer is passed through a sigmoid function. The predicted label is 1 if the value is equal to or above 0.5, otherwise the inferred label is 0.

3.6 Q5: Describe the chosen optimization method

The optimization method used in this fine-tuning task is *AdamW*, a variant of the standard *Adam* optimizer but with decoupled weight decay regularization [7]. While the standard *Adam* optimizer uses L_2 regularization directly to its gradient updates, the *AdamW* optimizer subtracts a scaled version of the weights in when updating the weights in the end of an iteration. This prevents interactions between weight decay and momentum vectors .

The optimization update rule for a parameter θ_t at time step t is given by

$$\theta_t = \theta_{t-1} - \eta_t \left(\frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} + \lambda \theta_{t-1} \right) \quad (5)$$

where \hat{m}_t and \hat{v}_t are the first and second moment estimates, η_t is the learning rate, λ is the weight decay coefficient, and ϵ is a small constant for numerical stability. A more detailed description of all updates in the optimization loop can be found in the paper *DECOUPLED WEIGHT DECAY REGULARIZATION* [7]. The adaptive learning rate using momentum vectors gives each parameter a learning rate based on past gradients. This can prevent getting stuck in local minima, since a small gradient leads to a smaller \hat{v}_t , and additionally stabilize when there are large gradients.

3.7 Comparison of BERT and perceptron classifier results

In table 2 below, we see the performance metrics-accuracy, precision, recall, and F1-score-of both the BERT model and the perceptron classifier when evaluated on the test set (25,000 movie reviews).

Model	Accuracy	Precision	Recall	F1-score
BERT-test	0.8212	0.8373	0.7973	0.8168
Perceptron-test	0.8702	0.8764	0.8620	0.8691

Table 2: Performance metrics comparison between BERT and Perceptron classifiers.

We find that the simple perceptron classifier achieves higher accuracy (0.8702 vs. 0.8212) than the complex BERT model. It also outperforms BERT in all other metrics precision, recall, and F1-score. This suggests that, for this particular dataset, the perceptron classifier is more effective at distinguishing between positive and negative reviews.

There could be several possible explanations for this, for example the following:

- **Perceptron Simplicity:** The perceptron classifier assigns a weight to each token in the vocabulary and updates the weights using the perceptron learning rule. The perceptron only updates weights when a misclassification occurs, which could be a factor in making it well-suited for linearly separable problems.
- **BERTs Need for Large-Scale Fine-Tuning:** BERT is a powerful model designed to capture deep contextual meaning, but fine-tuning on only 20,000 reviews may not be sufficient for it to fully generalize.
- **Dataset Characteristics:** Sentiment classification in movie reviews might be well captured by simpler language features instead of requiring the kind of deep contextual understanding that BERT provides.

A probable explanation is that it is due to several factors, and perhaps BERT is not able to use its strength on this data, while the perceptron classifier can utilize its strengths well. It should be noted that different versions of the BERT model might have a different performance.

4 Problem 2.4: Chatbot based on n-grams

In this problem, we build a chatbot based on n-grams, using Jelinek-Mercer smoothing. In this chapter, a brief user manual, implementation description and results will be presented.

4.1 User manual

In figure 3 below, we see the GUI of the implemented chatbot.

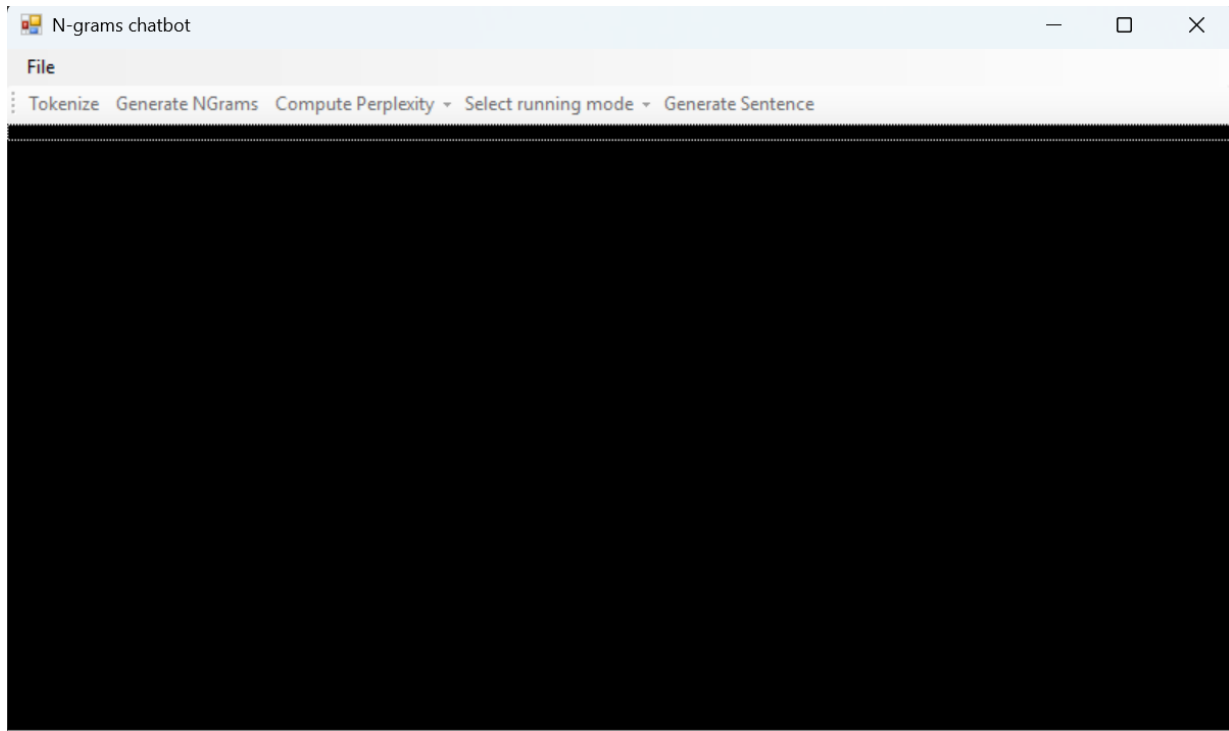


Figure 3: User interface of the implemented chatbot. It contains a button for loading each data set, tokenizing the data sets, generating n-grams, computing perplexity, a tool strip for selecting running mode for sentence generation and lastly a button for generating sentences in the selected mode.

In order to run the program, complete the following steps:

1. **Load data:**

Use the **File** button to load the training data, the validation data and the test data in this exact order. The buttons are disabled after the first click.

2. **Tokenize data:**

Press the **Tokenize** button, which tokenizes all three datasets. Enabled only once.

3. **Generate n-grams:**

Press the **Generate NGrams** button in order to generate the n-grams based on the training data set. Enabled only once.

4. **Compute perplexity:**

In order to compute perplexity, press the **Compute Perplexity** tool strip button and then press the button for the data set to compute over (training set, validation set or test set). Each button is disabled until the computation finishes.

Select running mode:

Use the **Select running mode** tool strip in order to select the running mode for generating sentences. The **Default mode** corresponds to *option (i)*: roulette wheel selection based on the probabilities of all tokens. The **Low-temperature mode** corresponds to *option (ii)*: roulette wheel selection over normalized probabilities for the top 10 most probable next tokens.

Generate sentence:

Use the **Generate Sentence** button in order to generate a sentence (which is generated in the selected running mode). This will generate a single sentence until an `jeosl` token, or up to a maximum of 100 tokens.

After completing the steps 1-3, the program is ready to run. Now in step 4, the user can either compute perplexities, or generate sentences. It is possible to do both at the same time, and compute several perplexities at the same time, but automatic clearing of the list box can cause the wanted information to disappear. Therefore, it is recommended to do one at a time.

4.2 Implementation description

In this section, the more important parts of the implementation will be described briefly.

4.2.1 Tokenization

Since the data is already preprocessed, the tokenization is kept simple. It's handled by the `Tokenizer` class, which splits sentences on newlines and then splits sentences into tokens on `"\s+"`-spaces. It returns the tokenized data set in `List<List<string>>` data structures, with the inner lists consisting of a full sentence.

4.2.2 N-grams

The handling of n-grams is built upon the solution from the n-grams autocompleter in problem 1.4, so it will be described briefly here.

The n-grams are handled by the `NGramsManager` class with it's main method being `GenerateNGrams`. This method, along with the helper method `CreateNGramDictionary`, creates `Dictionary<string, List<NGram>>` data structures for unigrams, bigrams and trigrams. The lookup key for bigrams is the previous word in the sentence, and for the trigrams the previous two words in the sentence, and for unigrams it is the "current" word. The `List<NGram>` structures contain all the `NGrams` objects which begin with the lookup key, and their corresponding frequency per million instances.

These n-gram dictionaries are used for example when generating a sentence or calculating probabilities, by accessing the list of all `<NGrams>` containing the last one or two tokens in the sentence. The n-grams are generated by pressing the `Generate NGrams` button, which builds the n-gram dictionaries using the tokenized training data, keeping track of each n-grams frequency.

4.2.3 Chatbot

The `Chatbot` class contains the main functionality of the n-grams based chatbot. In this subsection, its three main methods `JelinekMercerSmoothing`, `ComputePerplexity` and `GenerateSentence` will be described briefly.

JelinekMercerSmoothing:

This method takes a list of strings containing the previous words (of a sentence), and gives as output a `Dictionary<string, double>` structure containing each word and it's probability. The probability is based off of equations (6) and (7)

$$P_{JM}(t_k|t_{k-2}, t_{k-1}) = \alpha P_3(t_k|t_{k-1}, t_{k-2}) + \beta P_2(t_k|t_{k-1}) + (1 - \alpha - \beta) P_1(t_k) \quad (6)$$

$$P_{JM}(t_k|t_{k-1}, t_{k-1}) = \gamma P_2(t_k|t_{k-1}) + (1 - \gamma) P_1(t_k) \quad (7)$$

and is generated using the counts from the n-gram dictionaries. It takes as input previous tokens, for example the previous words during sentence generation. The probability distribution is sampled with a roulette wheel selection in the `SampleFromDistribution` helper method.

The `JelinekMercerSmoothing` method takes the the total frequencies of n-grams, and later stores the frequency of a specific n-gram divided by the total frequency. This with the lookup key being the next word, i.e. the next predicted word when generating a sentence.

This method works very well for sentence generation, since the lookup keys for the n-gram dictionaries contain the previous context making it quite fast. However it has a big flaw in being very slow when using it for computing perplexity. This because it loops over **all** possible next tokens giving a probability, instead of giving the probability for a specific next token. So, when computing probability for the next token in the perplexity calculation, we compute a lot of probabilities that are not used. I tried to fix this when i noted what the issue was, but unfortunately I realized this far too late and my attempts failed. Note however that the perplexity computation while slow gives values in the expected range, so I have no reason to believe the calculations are wrong.

ComputePerplexity:

As previously mentioned, due to a lot of unnecessary probability-calculations. It takes a tokenized data set `List<List<string>>` as input, and returns a `double`. It loops over all the sentences, computing probabilities with the `JelinekMercer` method. Then, it calculates the log probabilities in order to calculate the probabilities as equation (8) below:

$$PP(T) = \exp \left(-\frac{1}{n} (\log P(t_n|t_1, \dots, t_{n-1}) + \dots + \log P(t_2|t_1) + \log P(t_1)) \right) \quad (8)$$

This as to not get underflow issues, and taking the final exponent at the end, before returning the computed *perplexity*. This function, since it is quite slow, has a progress update to the GUI, using an `Action<int>` delegate. The time taken to run it, at least on my laptop, was around 45 minutes for the training data set, and around 8 minutes for the validation and test data sets.

GenerateSentence:

This method contains the functionality for generating a sentence using the Jelinek-Mercer smoothed probabilities. It takes a `bool lowTemperatureMode` as input, which decides if it should generate tokens using the *default mode* sampling from all probabilities, or the *low-temperature mode* sampling from the (normalized) top 10 most probable words. Additionally, it has a maximum amount of 100 generated tokens. Otherwise, it generates until the `<eos>` token is generated, finishing the sentence. It returns the full sentence (i.e. all tokens) as a string, separating the tokens with a space.

4.2.4 MainForm

There is not much needed to be said about the `MainForm` code, since it's functionality is described in the *User manual* subsection. It handles slower parts of the program using `Task.Run (() => ...` or separate specified threads as to maintain GUI operation. It handles GUI updates from non-main threads using the `ShowSafeProgress` method to ensure thread-safe operation.

4.3 Results

The model was tested by computing perplexity over the training data set, validation data set and test data set. First, with the initial parameter set ($\alpha = 0.25$, $\beta = 0.5$, $\gamma = 0.5$), and then with a

final parameter set ($\alpha = 0.3$, $\beta = 0.45$, $\gamma = 0.55$) found by manual testing on the validation set. The results are seen in the table 3 below.

Data Set	Initial Parameters	Final Parameters
	$(\alpha = 0.25, \beta = 0.5, \gamma = 0.5)$	$(\alpha = 0.3, \beta = 0.45, \gamma = 0.55)$
Training Set	19.47	17.62
Validation Set	190.55	189.69
Test Set	176.85	176.05

Table 3: Perplexity results for the initial and final parameter settings.

We can see a slight improvement with the new parameters, with the largest improvement being on the training data set. However, when testing around, most of the tried combination lead to worse results than the initial parameters.

In the textboxes below, we can see ten examples of generated text in each mode.

Default mode sentences

Sentence 1:

<bos> international business machines and a career by \$ n cents a share a n billion in back taxes those fed to disclose to and design <eos>

Sentence 2:

<bos> the inflation rate manufacturing the bills will take over recognition the <unk> a rising prices <eos>

Sentence 3:

<bos> the sec 's the current that transfer be and earnings <eos>

Sentence 4:

<bos> the honda purchase reinsurance but for the national statistics office <unk> was previously bid marker clark does n't support it <eos>

Sentence 5:

<bos> for mr. utilities underwriters led that when \$ n a 've had now president hire more people n years ago <eos>

Sentence 6:

<bos> organized labor says <unk> weeks and company head of this group a yield of marketing profits and everyone approach did n't <eos>

Sentence 7:

<bos> however less developed the a reasonably attractive proposition of <unk> sold more than twice has <unk> up from win <eos>

Sentence 8:

<bos> this change <eos>

Sentence 9:

<bos> years in prison and <eos>

Sentence 10:

<bos> monday about the concept that <eos>

Low-temperature mode sentences

Sentence 1:

<bos> in fact <eos>

Sentence 2:

<bos> the <unk> a <unk> <unk> <eos>

Sentence 3:

<bos> the first to agree to the u.s. the hartford where it 's a share of the market <unk> <eos>

Sentence 4:

<bos> <unk> a new and the new york stock <unk> n <eos>

Sentence 5:

<bos> <unk> said <eos>

Sentence 6:

<bos> <unk> to <unk> to \$ n million <eos>

Sentence 7:

<bos> <unk> of <unk> n million for <unk> <unk> <eos>

Sentence 8:

<bos> but the chicago mercantile exchange <eos>

Sentence 9:

<bos> but <unk> of the the <unk> <unk> in the third quarter <eos>

Sentence 10:

<bos> in the <unk> <unk> the <unk> n n million or \$ n to million in the <unk> to n <eos>

I would say from testing around, that there is more variation and longer sentences in general from the default mode, and more <unk> tokens in the low-temperature mode. However, neither of them generate meaningful sentences, as is expected with the small n-grams used here. The low temperature mode seems more repetitive, and more likely to generate the <eos>, which is to be expected since it leverages the most common n-grams. Since the default mode generates more actual words, instead of unknown words, it's sentences can look cleaner and closer to a "real" sentence.

References

- [1] SMHI, "Sveriges sjötätaste landskap." <https://www.smhi.se/kunskapsbanken/hydrologi/sveriges-sjoar/sveriges-sjotataste-landskap->. Accessed: 2025-03-06.
- [2] T. Bull, "The sámí language(s), maintenance and intellectualisation," *Current Issues in Language Planning*, vol. 3, no. 1, pp. 28–39, 2002.

- [3] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, “Well-read students learn better: On the importance of pre-training compact models,” 2019.
- [4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018.
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023.
- [6] Keras, “Dropout layer.” https://keras.io/api/layers/regularization_layers/dropout/. Accessed: 2025-03-09.
- [7] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2019.