INTERACTION WITH LARGE SCALE MODELS

LIACD/1

University of Beira Interior,
Department of Informatics

Hugo Pedro Proença,
hugomcp@di.ubi.pt, 2024/2025

INTERACTION WITH LARGE SCALE MODELS

[06]

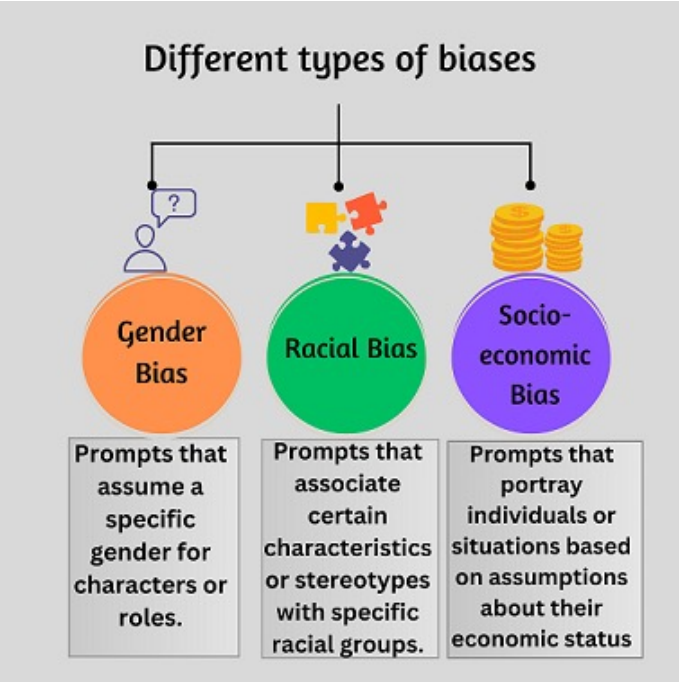
- Common pitfalls in prompt engineering: ambiguity, bias, and verbosity.
- Troubleshooting inconsistent or undesirable outputs.
- Debugging techniques: Simplification, incremental changes, and error analysis.

Prompt Bias

- Prompt-induced bias can affect tone, inclusivity, and neutrality.
  - Example: 'Why is remote work worse than in-office work?' ❌
  - Fix: 'Compare the pros and cons of remote vs. in-office work.' ✅
- P**rompt bias** refers to the unintended influence that the wording, structure, or context of a prompt can have on the output of a language model. Because LLMs are highly sensitive to input phrasing, even small changes in the prompt can lead to significantly different responses in tone, content, or correctness.
  - Introduce subjective framing (e.g., emotionally loaded language).
  - Steer models toward particular worldviews.
  - Affect fairness, factuality, or diversity of outputs.
  - Be exploited intentionally (manipulative prompts).
  - Occur unintentionally due to prompt design.

Prompt Bias

- Why Prompt Bias Matters
  - Ethical concerns: Biased prompts can reinforce stereotypes or marginalize certain groups.
  - Reproducibility issues: Slight changes in prompt phrasing may lead to inconsistent results.
  - Evaluation distortion: Prompt bias can inflate or deflate model performance in benchmarks.
  - Scientific integrity: Bias skews data generation and analysis when using LLMs in research or content creation.



Source: https://medium.com/@simrwaraich/

Prompt Bias

Ensuring Unbiased Responses in Political or Social Topics

LLMs may generate politically biased or opinionated responses if the prompt is vague.

Weak Prompt:

"What is the best political system in the world?" 

This can lead to **subjective, biased responses** based on model pretraining data.

Improved Prompt:

"Provide an objective comparison of different political systems (e.g., democracy, autocracy, monarchy) based on key factors such as governance efficiency, economic impact, and civil liberties. Use data from reputable sources, and do not express personal opinions."

- **Encourages neutrality** by requesting an objective comparison.
- **Minimizes bias** by requiring key evaluation factors.
- **Reduces hallucinations** by instructing the model to rely on **reputable sources**.

Prompt Bias – Examples

- **Leading Question Bias**
  - *"Why is solar power better than nuclear power?"*
    - This assumes a preferred answer and steers the model toward supporting solar power.
  - **Better Prompt:**
  - "Compare the advantages and disadvantages of solar power and nuclear power."

- **Framing Bias**
  - "As a conservative politician, explain why universal healthcare is problematic."
    - This restricts the model's persona and primes a partisan view.
  - **Better Prompt:**
  - "Explain the arguments for and against universal healthcare from different political perspectives."

Prompt Bias – Examples

- **Sentiment Bias**
  - *"Describe the failures of remote education during the pandemic."* ✗
    - The model will likely focus on negative outcomes.
  - **Better Prompt:**
  - ""Describe the outcomes of remote education during the pandemic, including both benefits and challenges."

- **Cultural Bias**
  - "What is the typical family structure?" ✗
    - LLMs might assume a Western nuclear family structure.
  - **Better Prompt:**
  - "Describe common family structures across different cultures."

Prompt Bias – Examples

- **Priming with Context**
  - Roles might be a primary factor of bias. It is up to the human to decide whether such biases should be taken into account or not.
  - Most probably, the responses coming from “specific role characters” should be taken with salt-and-pepper.

  - *"You are an expert scientist. What are the causes of climate change?”*

  - *“You are a factory owner. What are the causes of climate change?”*

- *These two identities may cause the model to focus on different causes or tone due to role-based bias.*

Prompt Bias – Detect and Minimize

- A/B Testing
  - Try different phrasings and compare outputs for tone, coverage, and fairness.
- Diversify Roles and Perspectives
  - Use prompt variations to expose the model to multiple viewpoints.
- Prompt Auditing
  - Ask the model to critique its own output for bias.
- Use Meta-Instructions
  - e.g., "Give a neutral and balanced overview...".
- Avoid Loaded Language
  - Rephrase prompts to remove emotionally or ideologically charged words.

Prompting with intention: Craft prompts with awareness of potential biases and strive for inclusivity and fairness in language choices.

Critical evaluation: Don't blindly accept AI outputs as truth; critically evaluate them, verify information, and identify potential biases.

Responsible use: Be mindful of the potential risks associated with AI outputs and avoid using them for malicious purposes.

Advocacy for ethical AI: Support initiatives promoting transparency, accountability, and responsible development practices in the field of AI.

Prompt Ambiguity

- Prompt Ambiguity arises when a prompt is unclear, vague, or open to multiple interpretations. This can lead to inconsistent or irrelevant responses from an LLM.
  - Ambiguity can result from:
  - Vague or underspecified instructions
  - Grammatical ambiguity
  - Polysemous (multiple-meaning) words
  - Unclear referents or context
  - Implicit assumptions about prior knowledge
- Why Prompt Ambiguity Matters
  - Reduces reproducibility: Same prompt may yield different answers across runs or model versions.
  - Decreases accuracy: Model might "guess" what the user meant and go in the wrong direction.
  - Wastes resources: More iterations needed to refine the prompt or validate outputs.
  - Harms user trust: Unreliable outputs reduce usability in real-world systems (e.g., chatbots, research assistants, decision support tools).

Prompt Ambiguity - Lexical

- Prompt:

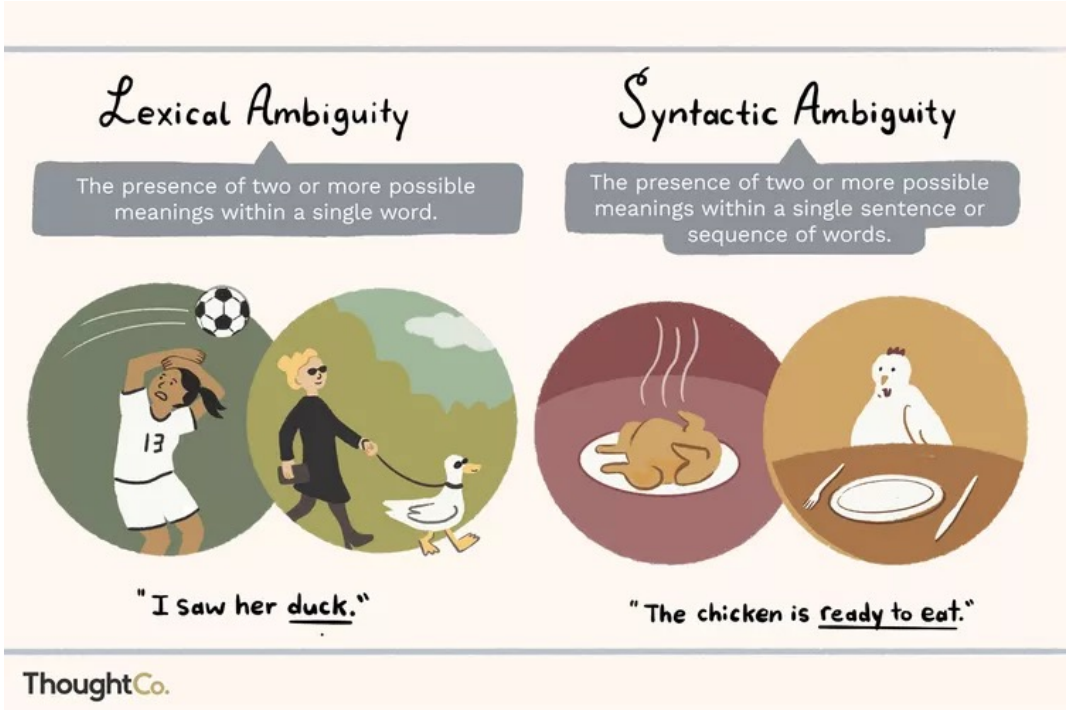
  "*Describe a bat.*"

- In this case, the word bat can mean:
  - A flying mammal
  - A baseball/cricket tool

- Better Prompts:

  "*Describe a bat, the animal.*"

  "*Describe a bat used in baseball.*"



Source: https://www.thoughtco.com/ambiguity-language-1692388

Prompt Ambiguity - Referential

- Prompt:

"*John met Tom at the café.*

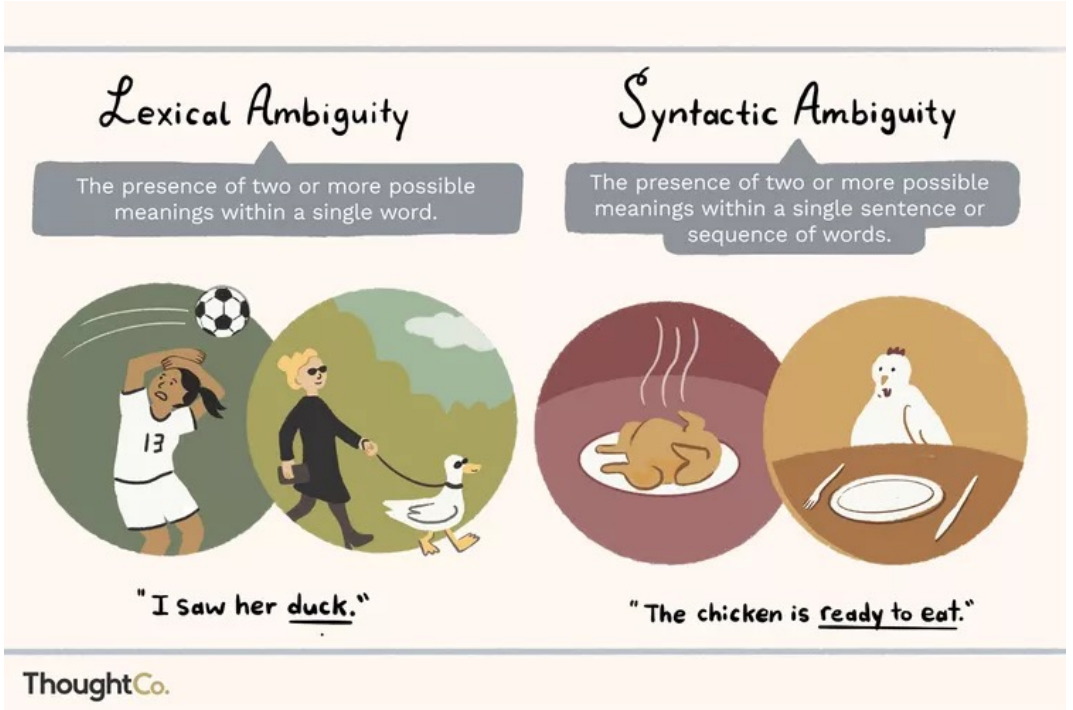
He was very nervous. Why?."

- Who was nervous — John or Tom?

- Better Prompts:

"*John met Tom at the café.*

John was very nervous. Why?."



Source: https://www.thoughtco.com/ambiguity-language-1692388

Prompt Scope and Instructional Ambiguity

- **Scope Ambiguity**

  "*List the students in the AI course who passed and failed."*

- Are we listing:
  - All students and indicating who passed/failed?
  - Only those who passed and those who failed (excluding others)?
  - Or is it asking for two separate lists?
- Better Prompt:

"*Provide two separate lists: one of students who passed the AI course, and another of those who failed."*

- ***Instructional Ambiguity***

  "Summarize the article and provide feedback."

- *Unclear:*
  - *How long should the summary be?*
  - *What kind of feedback — stylistic, factual, tone, etc.?*
- *Better Prompt:*

"Write a 3-sentence summary of the article, then provide feedback on its tone and clarity."

Contextual Ambiguity

- Consider the following prompt (Romeo and Juliet, beginning).

"A street fight breaks out between the Montagues and the Capulets, which is broken up by the ruler of Verona, Prince Escalus. He threatens the Montagues and Capulets with death if they fight again.

A melancholy Romeo enters and is questioned by his cousin Benvolio, who learns that the cause of Romeo's sadness is unrequited love."

Continue this story"

- Without the preceding context or clear tone, the model may take the story in any direction (comedy, horror, drama...)

- **Better Prompt:**
- *"Continue the romantic story set in such a way that at the end the main characters stay together and have a long happy life…"*

Prompt Debugging Example

🧰 **Goal:**

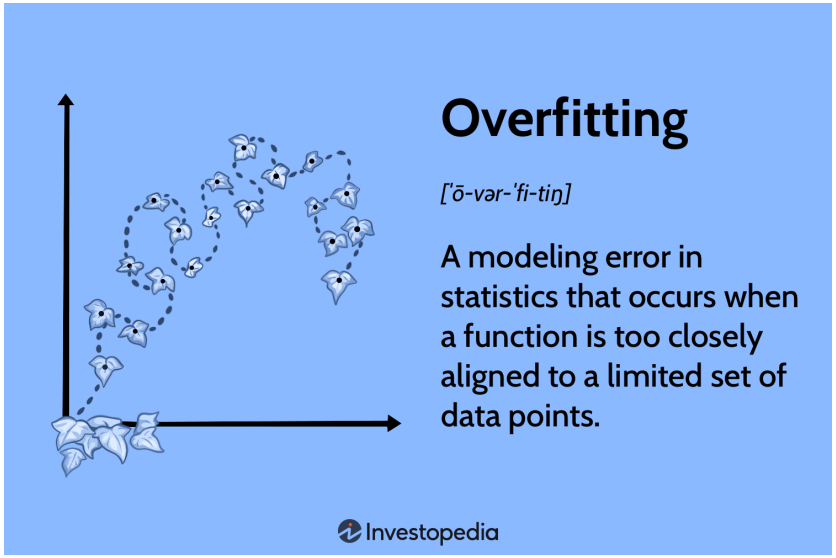
- Create a prompt that makes an LLM explain **overfitting** to a **non-technical audience** in a **concise and clear** way.

- Step 1: We write the initial prompt

  "*Explain Overfitting*"

🔍 **Output Problems:**

- Might be too technical.
- Uses jargon like "*variance*," "*model complexity*" or "*generalization error*."
- No context about audience or length.



Source: https://https://www.investopedia.com/terms/o/overfitting.asp

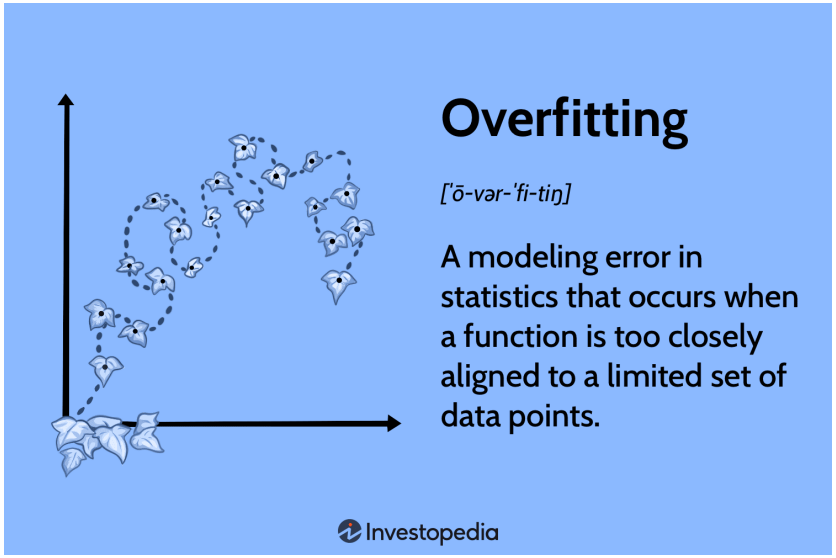
Prompt Debugging Example

🧰 **In a second step, we add some format hints**

"*Explain Overfitting in simple terms*"

🔍 **Output Problems:**

- Slightly better.
- Still uses phrases like "test data" and "training data" without explanation.
- No analogies or simplification for laypeople.



Source: https://https://www.investopedia.com/terms/o/overfitting.asp

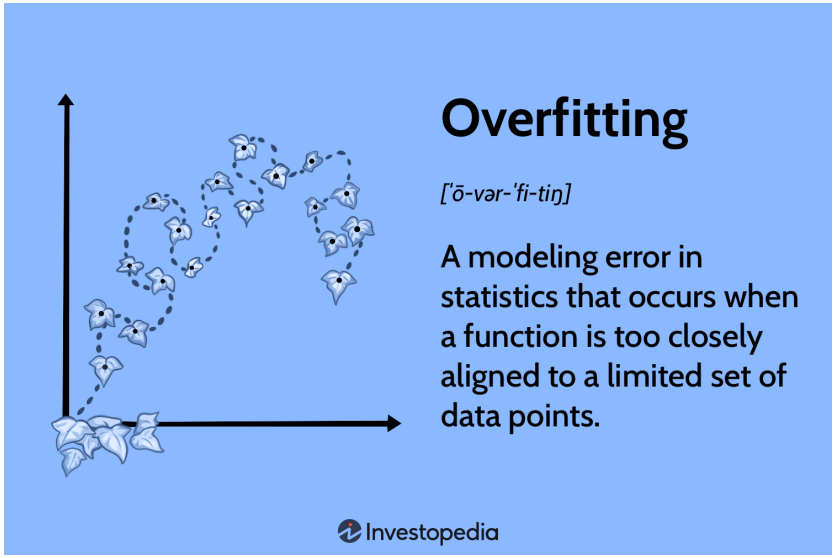
Prompt Debugging Example

🧰 **Next, we specify the Audience**

"*Explain overfitting in simple terms to someone with no background in machine learning.*"

🔍 **Output Problems:**

- Better explanation.
- Still might be a bit dry or abstract.
- Could benefit from an **analogy** or **real-life example**.

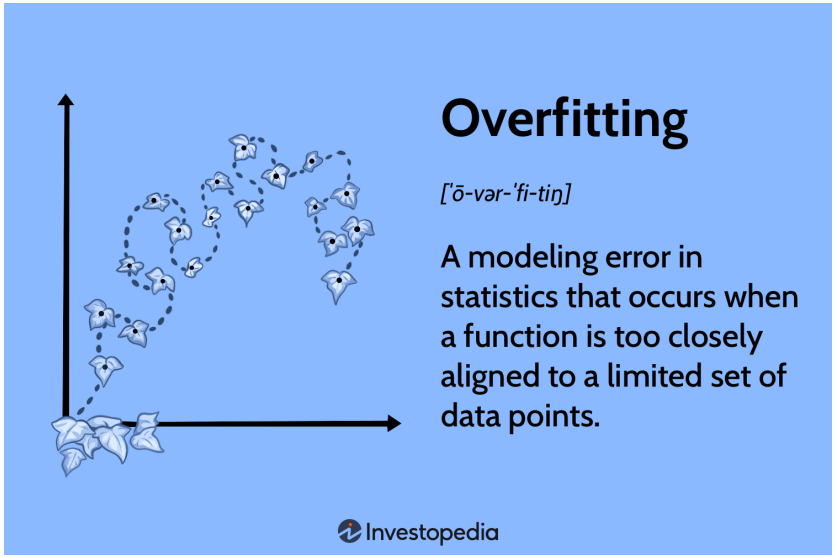


Source: https://https://www.investopedia.com/terms/o/overfitting.asp

Prompt Debugging Example

🧰 **Next, Add a Metaphor or Analogy Instruction**

"*Explain overfitting in simple terms to someone with no background in machine learning. Use a real-life analogy.*



Source: https://https://www.investopedia.com/terms/o/overfitting.asp

✅ **Output:**

- "*Overfitting is like a student who memorizes practice questions for an exam instead of understanding the material…*"
- Much clearer, engaging, and accessible.

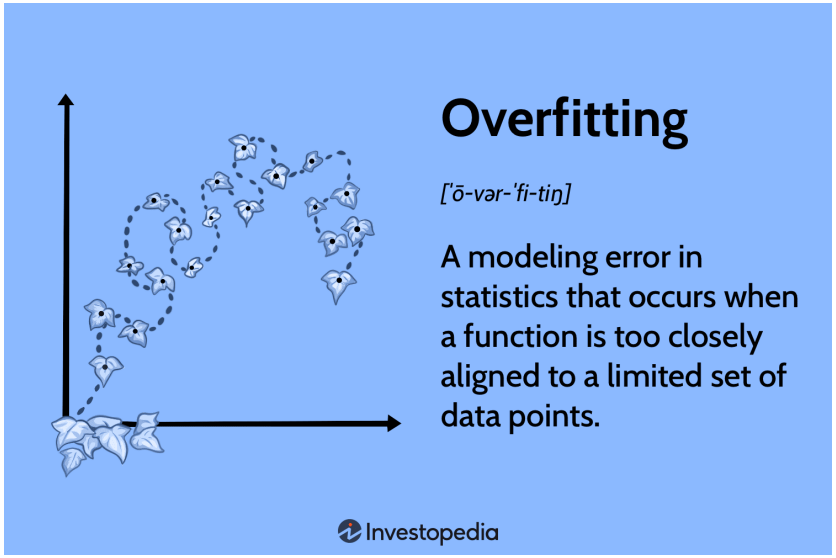
Prompt Debugging Example

🧰 **Finally, we add some output constraints (length, Tone, Style)**

"*Explain overfitting in simple terms to Explain overfitting in simple terms to someone with no background in machine learning. Use a real-life analogy. Keep the explanation under 100 words and make it friendly in tone.*"

✅ **Final Output:**

- *"Overfitting is like a student who memorizes answers to practice tests but struggles with the real exam. The student did great on the questions they saw before but didn't truly understand the material. In machine learning, a model that overfits learns the training data too well and fails to perform on new, unseen data."*



Source: https://www.investopedia.com/terms/o/overfitting.asp

Prompt Debugging Summary and Tool

Criterion	Check
Who is the audience?	✅ Specified clearly
What is the goal?	✅ Defined (explain clearly)
Is the tone correct?	✅ Friendly/simple/formal?
Any formatting needed?	✅ Word limit/structure
Is ambiguity reduced?	✅ Analogies/examples added

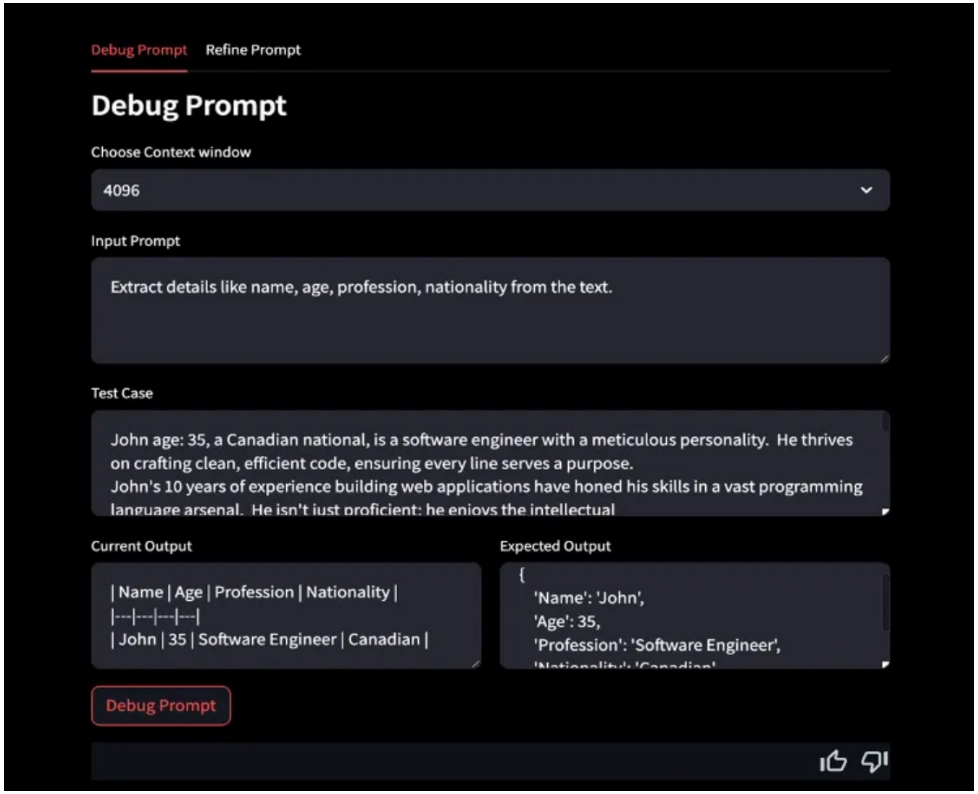
- A **Prompt Debugger** tool is available at (https://prompt-debugger-lbgzisv3qa-uc.a.run.app/)

and currently supports 2 functionalities: Debug Prompt and Refine Prompt.

1. **Debug Prompt** — Helps to debug prompt based on the set of instructions, the test case and the expected output.
2. **Refine Prompt** — rewrites any existing prompt based on the best practices of Prompt Engineering.

Prompt Debugger

1. **Context Window** — based on this selection the tool will restrict the token count. It will notify when your prompt exceeds the window size
2. **Input Prompt** — this is where we paste our instructions
3. **Test Case** — it takes care of the context we want to test the prompt on
4. **Current Output** — of all the fields this is the only non-mandatory field. As the name suggests you can share the current output that you are getting from your existing prompt.
5. **Expected Output** — a mandatory field where the expected output should be given



Source: https://prompt-debugger-lbgzisv3qa-uc.a.run.app/