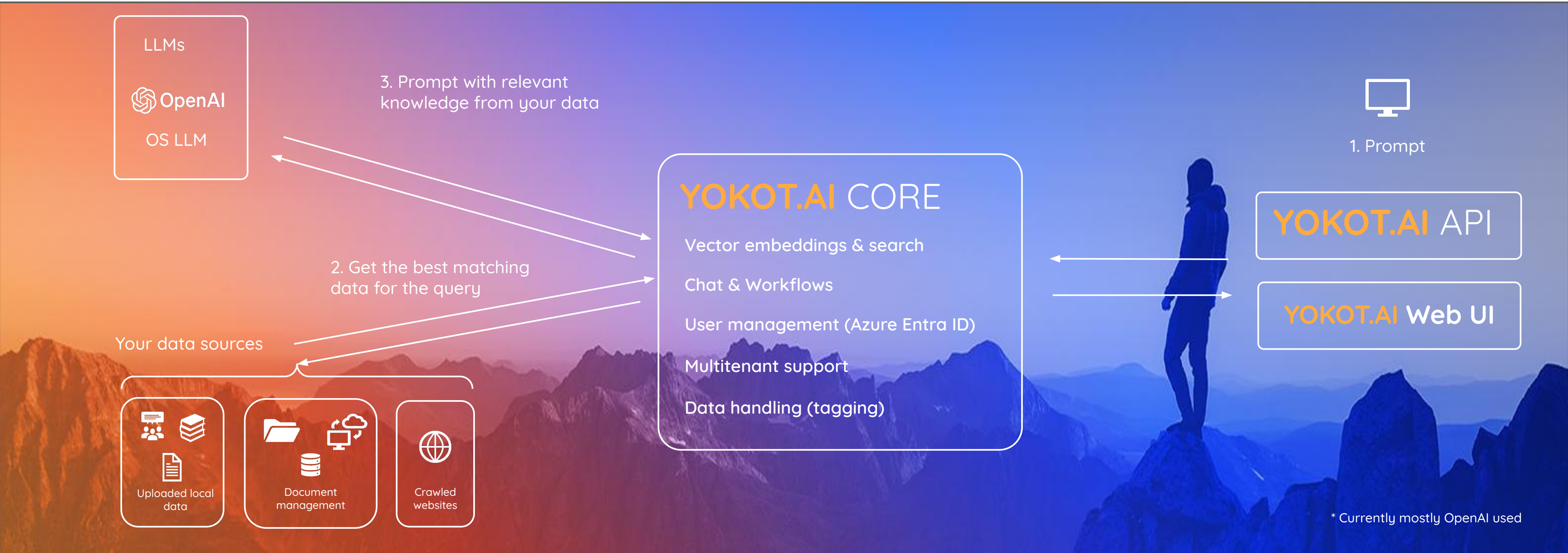
HOW TO TRUST AN LLM

YOKOT.AI case study

Mikko Lehtimäki, PhD, CTO, Softlandia

YOKOT.AI

More than RAG



LLMs

OpenAI

OS LLM

3. Prompt with relevant knowledge from your data

2. Get the best matching data for the query

Your data sources

- Uploaded local data
- Document management
- Crawled websites

YOKOT.AI CORE

- Vector embeddings & search
- Chat & Workflows
- User management (Azure Entra ID)
- Multitenant support
- Data handling (tagging)

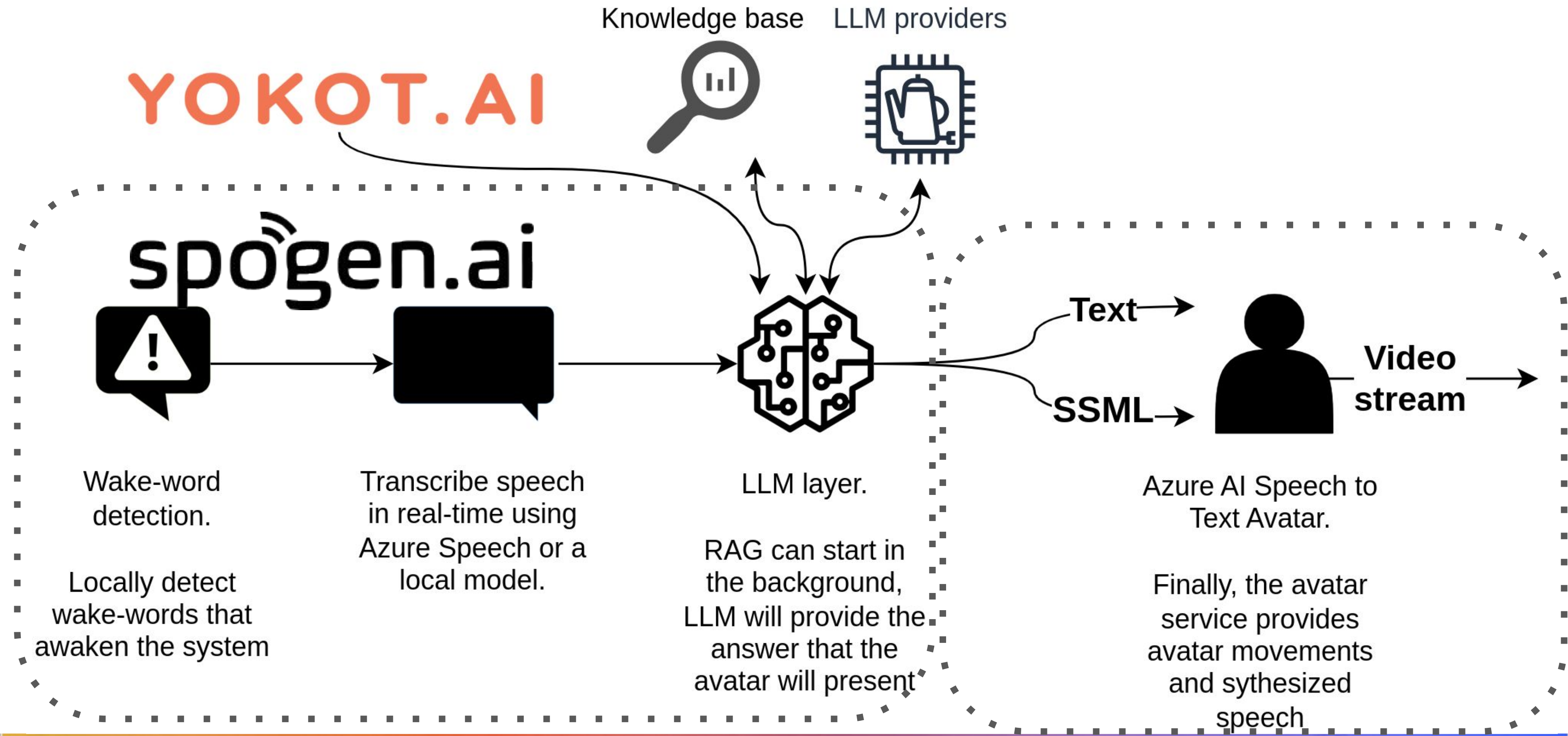
1. Prompt

YOKOT.AI API

YOKOT.AI Web UI

* Currently mostly OpenAI used

Technology



Wake-word detection.

Locally detect wake-words that awaken the system

Transcribe speech in real-time using Azure Speech or a local model.

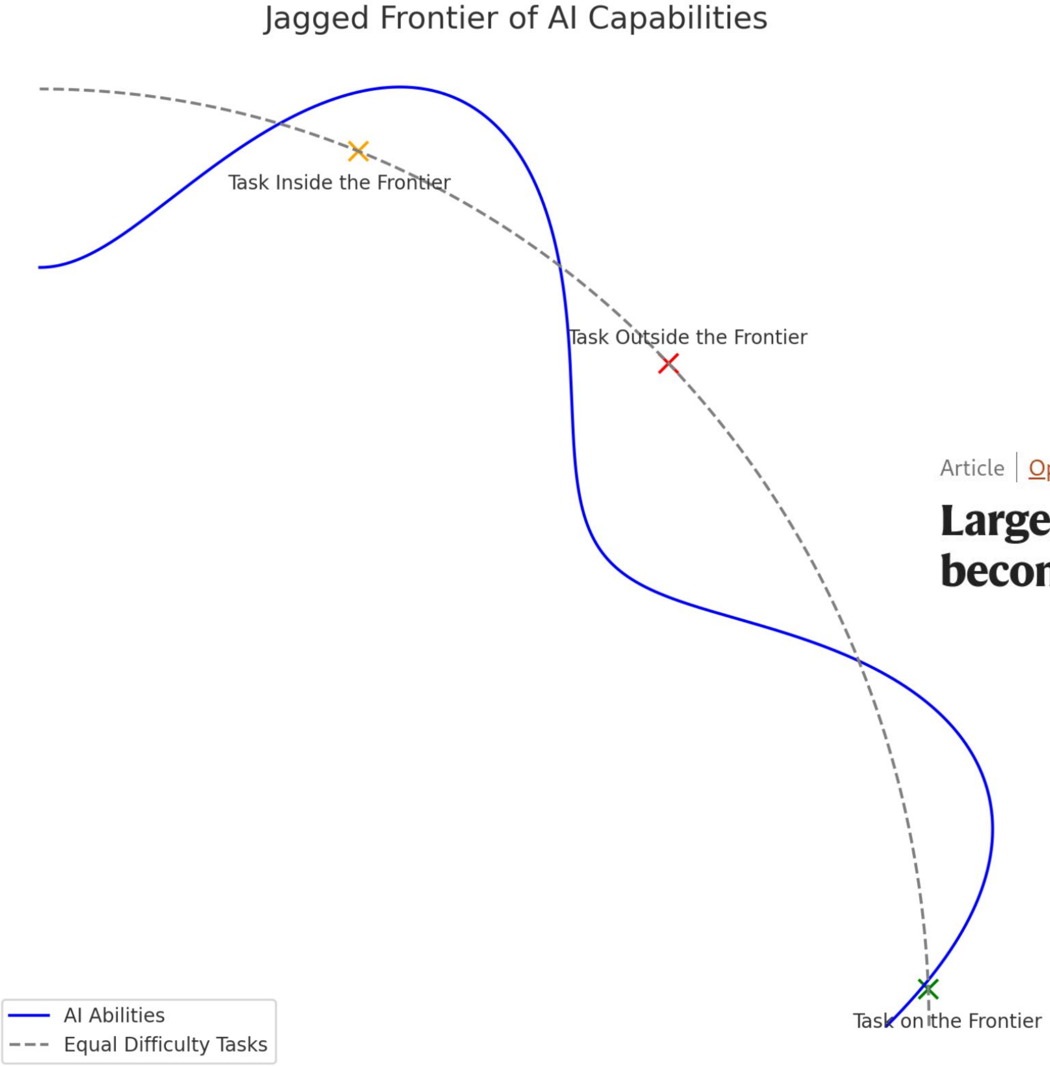
LLM layer.

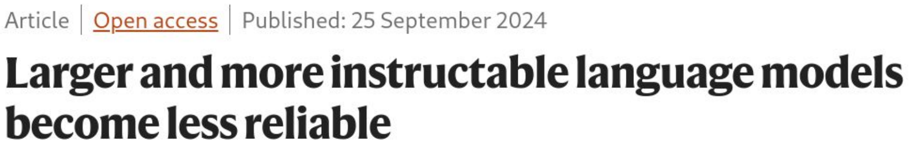
RAG can start in the background, LLM will provide the answer that the avatar will present

Azure AI Speech to Text Avatar.

Finally, the avatar service provides avatar movements and sythesized speech

LLM capabilities can surprise!




Article | Open access | Published: 25 September 2024

Larger and more instructable language models become less reliable

Dell'Acqua et. al., 2023, *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality,* Harvard Business School & Boston Consulting Group

How to develop a reliable LLM solution

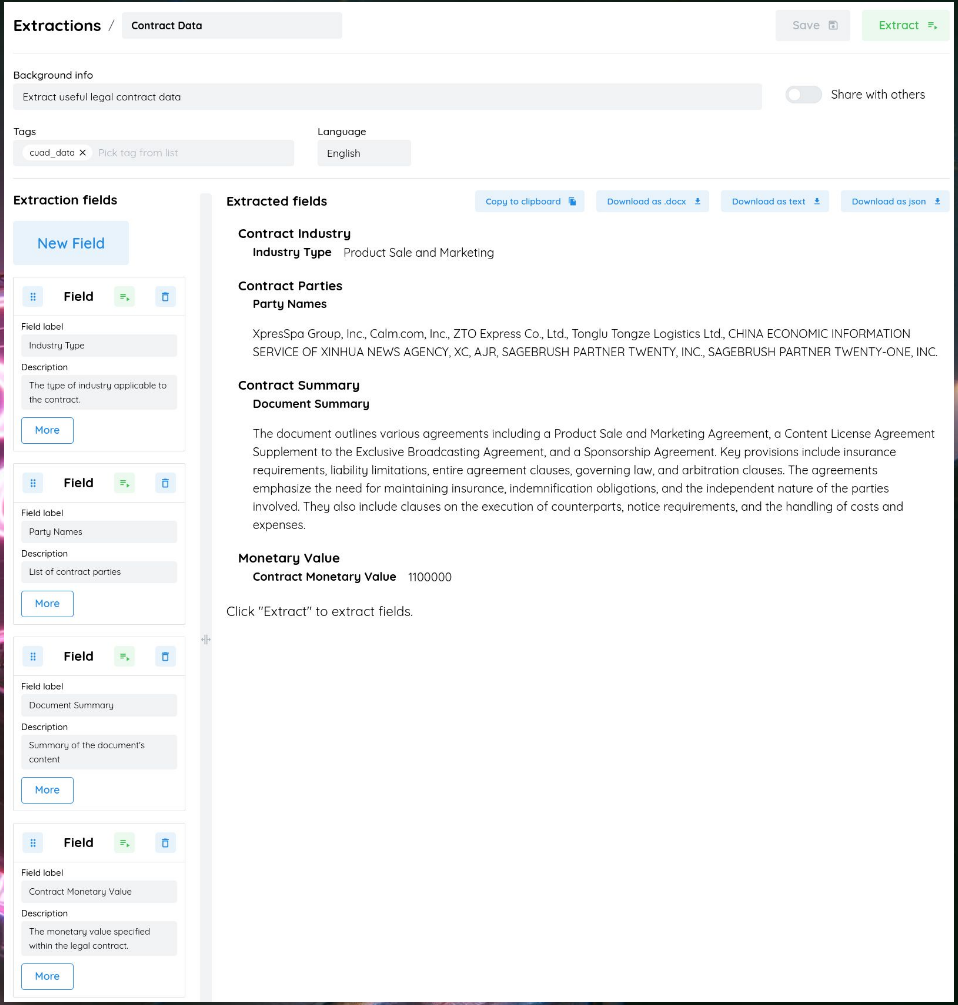
It’s a mix of traditional SW and ML.

- Define the task
- Observe the behavior
- Evaluate the answers

These steps are useful even when using an LLM as a development tool, and LLMs can help us along the way!

Case Study: structured data extraction

1. Determine data structure from user instructions
2. Extract structured data across a large number of documents



Extractions / Contract Data

Background info: Extract useful legal contract data

Tags: cuad_data

Language: English

Extraction fields

Field label: Industry Type
Description: The type of industry applicable to the contract.

Field label: Party Names
Description: List of contract parties

Field label: Document Summary
Description: Summary of the document's content

Field label: Contract Monetary Value
Description: The monetary value specified within the legal contract.

Extracted fields

Contract Industry

Industry Type Product Sale and Marketing

Contract Parties

Party Names

XpresSpa Group, Inc., Calm.com, Inc., ZTO Express Co., Ltd., Tonglu Tongze Logistics Ltd., CHINA ECONOMIC INFORMATION SERVICE OF XINHUA NEWS AGENCY, XC, AJR, SAGEBRUSH PARTNER TWENTY, INC., SAGEBRUSH PARTNER TWENTY-ONE, INC.

Contract Summary

Document Summary

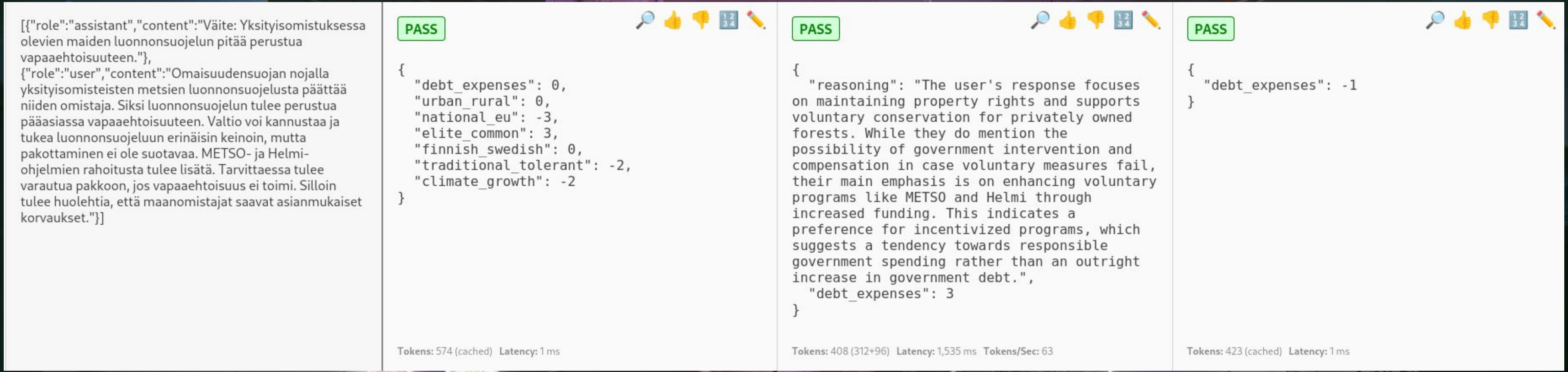
The document outlines various agreements including a Product Sale and Marketing Agreement, a Content License Agreement Supplement to the Exclusive Broadcasting Agreement, and a Sponsorship Agreement. Key provisions include insurance requirements, liability limitations, entire agreement clauses, governing law, and arbitration clauses. The agreements emphasize the need for maintaining insurance, indemnification obligations, and the independent nature of the parties involved. They also include clauses on the execution of counterparts, notice requirements, and the handling of costs and expenses.

Monetary Value

Contract Monetary Value 1100000

Click "Extract" to extract fields.

The problem



[{"role":"assistant","content":"Väite: Yksityisomistuksessa olevien maiden luonnonsuojelun pitää perustua vapaaehtoisuuteen."}, {"role":"user","content":"Omaisuudensuojan nojalla yksityisomisteisten metsien luonnonsuojelusta päättää niiden omistaja. Siksi luonnonsuojelun tulee perustua pääasiassa vapaaehtoisuuteen. Valtio voi kannustaa ja tukea luonnonsuojeluun erinäisin keinoin, mutta pakottaminen ei ole suotavaa. METSO- ja Helmi-ohjelmien rahoitusta tulee lisätä. Tarvittaessa tulee varautua pakkoon, jos vapaaehtoisuus ei toimi. Silloin tulee huolehtia, että maanomistajat saavat asianmukaiset korvaukset."}]

```
{
  "debt_expenses": 0,
  "urban_rural": 0,
  "national_eu": -3,
  "elite_common": 3,
  "finnish_swedish": 0,
  "traditional_tolerant": -2,
  "climate_growth": -2
}
```

Tokens: 574 (cached) **Latency:** 1 ms

```
{
  "reasoning": "The user's response focuses
on maintaining property rights and supports
voluntary conservation for privately owned
forests. While they do mention the
possibility of government intervention and
compensation in case voluntary measures fail,
their main emphasis is on enhancing voluntary
programs like METSO and Helmi through
increased funding. This indicates a
preference for incentivized programs, which
suggests a tendency towards responsible
government spending rather than an outright
increase in government debt.",
  "debt_expenses": 3
}
```

Tokens: 408 (312+96) **Latency:** 1,535 ms **Tokens/Sec:** 63

```
{
  "debt_expenses": -1
}
```

Tokens: 423 (cached) **Latency:** 1 ms

Output: 0

Output: 3

Output: -1

Same model, same task, slightly different prompts, different outcomes

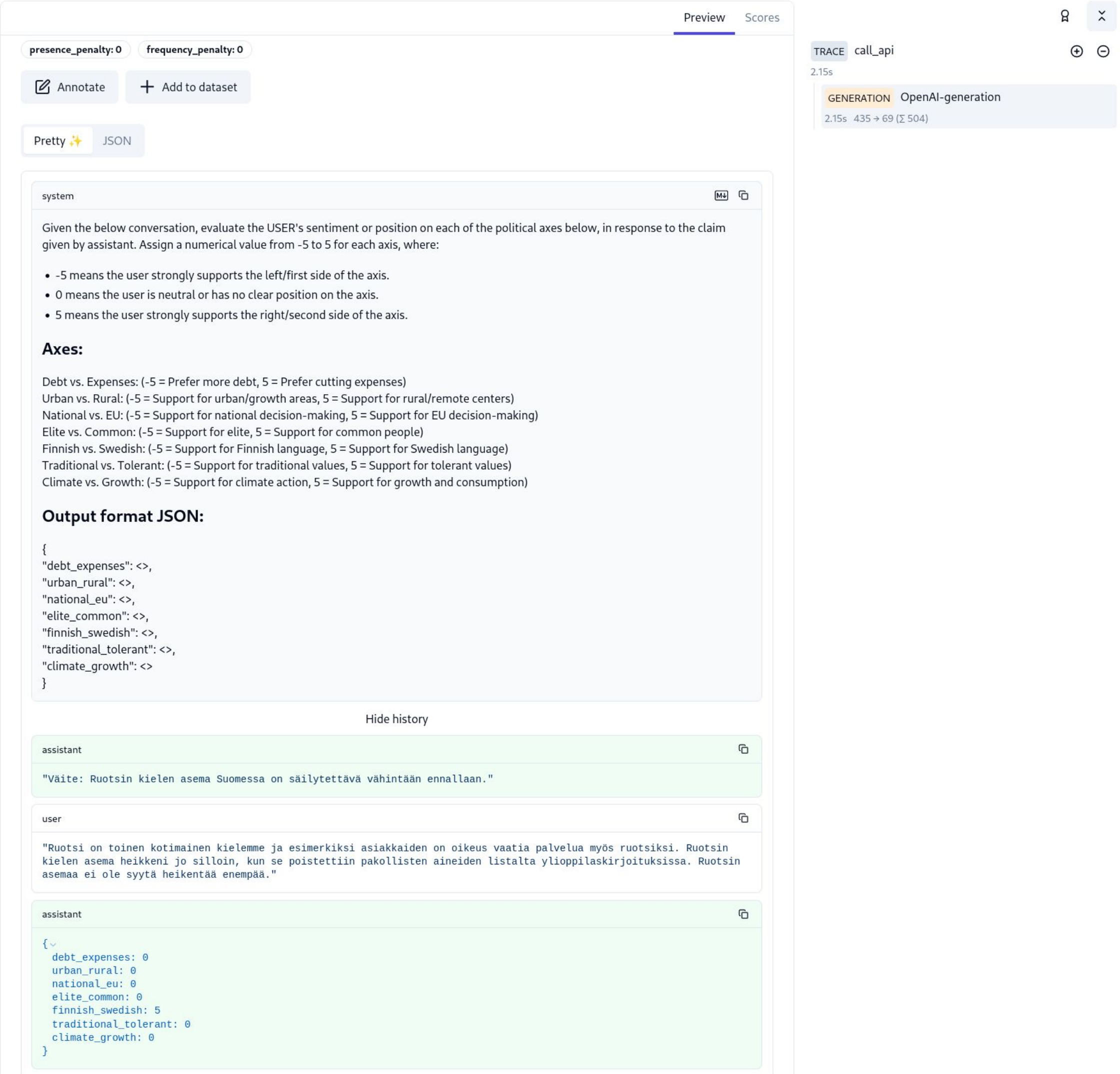
Software: Promptfoo

Observability

Precise prompts and outputs must be stored, tracked, evaluated!

Relevant during development and in production.

Software: Langfuse



presence_penalty: 0 frequency_penalty: 0

system

Given the below conversation, evaluate the USER's sentiment or position on each of the political axes below, in response to the claim given by assistant. Assign a numerical value from -5 to 5 for each axis, where:

- -5 means the user strongly supports the left/first side of the axis.
- 0 means the user is neutral or has no clear position on the axis.
- 5 means the user strongly supports the right/second side of the axis.

Axes:

Debt vs. Expenses: (-5 = Prefer more debt, 5 = Prefer cutting expenses)
Urban vs. Rural: (-5 = Support for urban/growth areas, 5 = Support for rural/remote centers)
National vs. EU: (-5 = Support for national decision-making, 5 = Support for EU decision-making)
Elite vs. Common: (-5 = Support for elite, 5 = Support for common people)
Finnish vs. Swedish: (-5 = Support for Finnish language, 5 = Support for Swedish language)
Traditional vs. Tolerant: (-5 = Support for traditional values, 5 = Support for tolerant values)
Climate vs. Growth: (-5 = Support for climate action, 5 = Support for growth and consumption)

Output format JSON:

```
{
"debt_expenses": <>,
"urban_rural": <>,
"national_eu": <>,
"elite_common": <>,
"finnish_swedish": <>,
"traditional_tolerant": <>,
"climate_growth": <>
}
```

assistant

```
"Väite: Ruotsin kielen asema Suomessa on säilytettävä vähintään ennallaan."
```

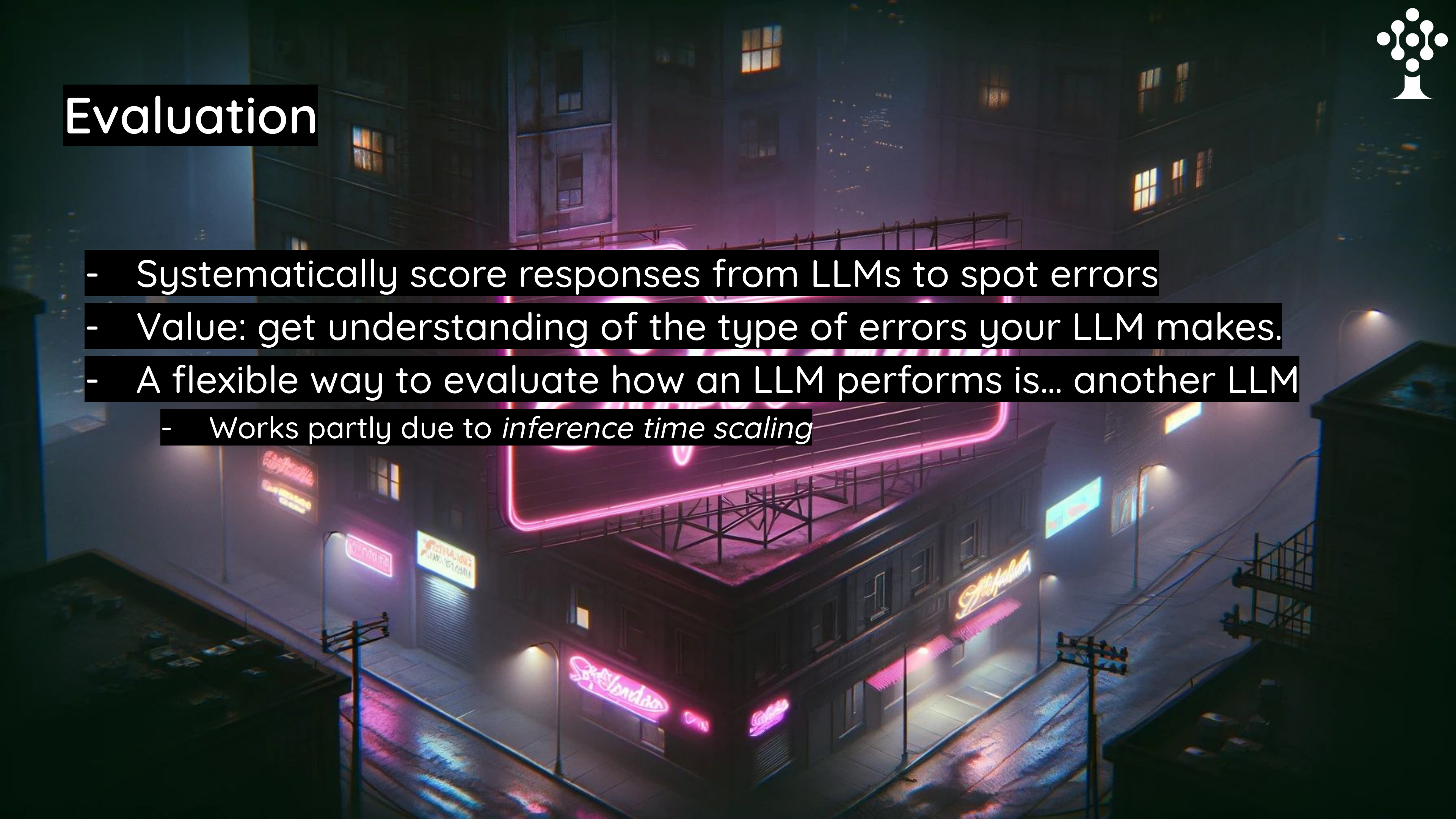
user

```
"Ruotsi on toinen kotimainen kielemme ja esimerkiksi asiakkaiden on oikeus vaatia palvelua myös ruotsiksi. Ruotsin
kielen asema heikkeni jo silloin, kun se poistettiin pakollisten aineiden listalta ylioppilaskirjoituksissa. Ruotsin
asemaa ei ole syytä heikentää enempää."
```

assistant

```
{
  debt_expenses: 0
  urban_rural: 0
  national_eu: 0
  elite_common: 0
  finnish_swedish: 5
  traditional_tolerant: 0
  climate_growth: 0
}
```

Evaluation

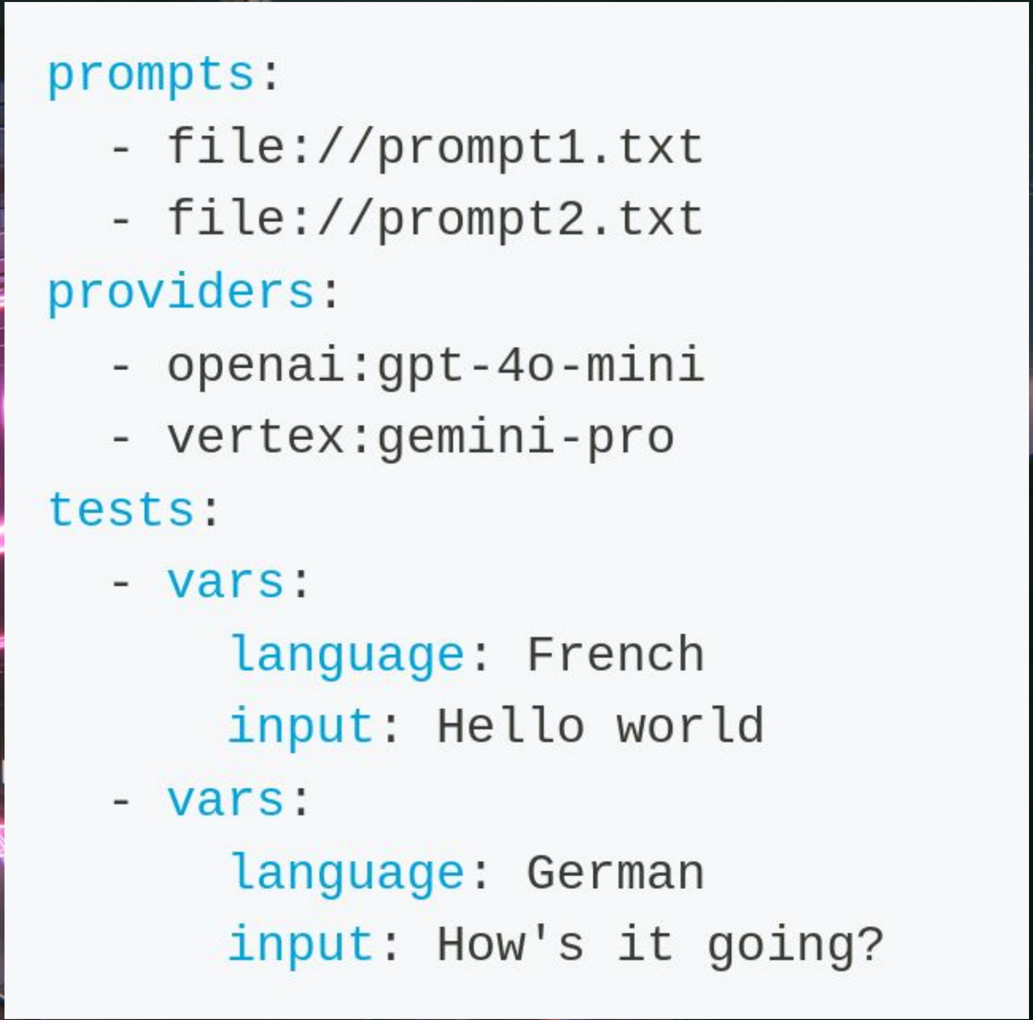


- Systematically score responses from LLMs to spot errors
- Value: get understanding of the type of errors your LLM makes.
- A flexible way to evaluate how an LLM performs is… another LLM
  - Works partly due to *inference time scaling*

Evaluation

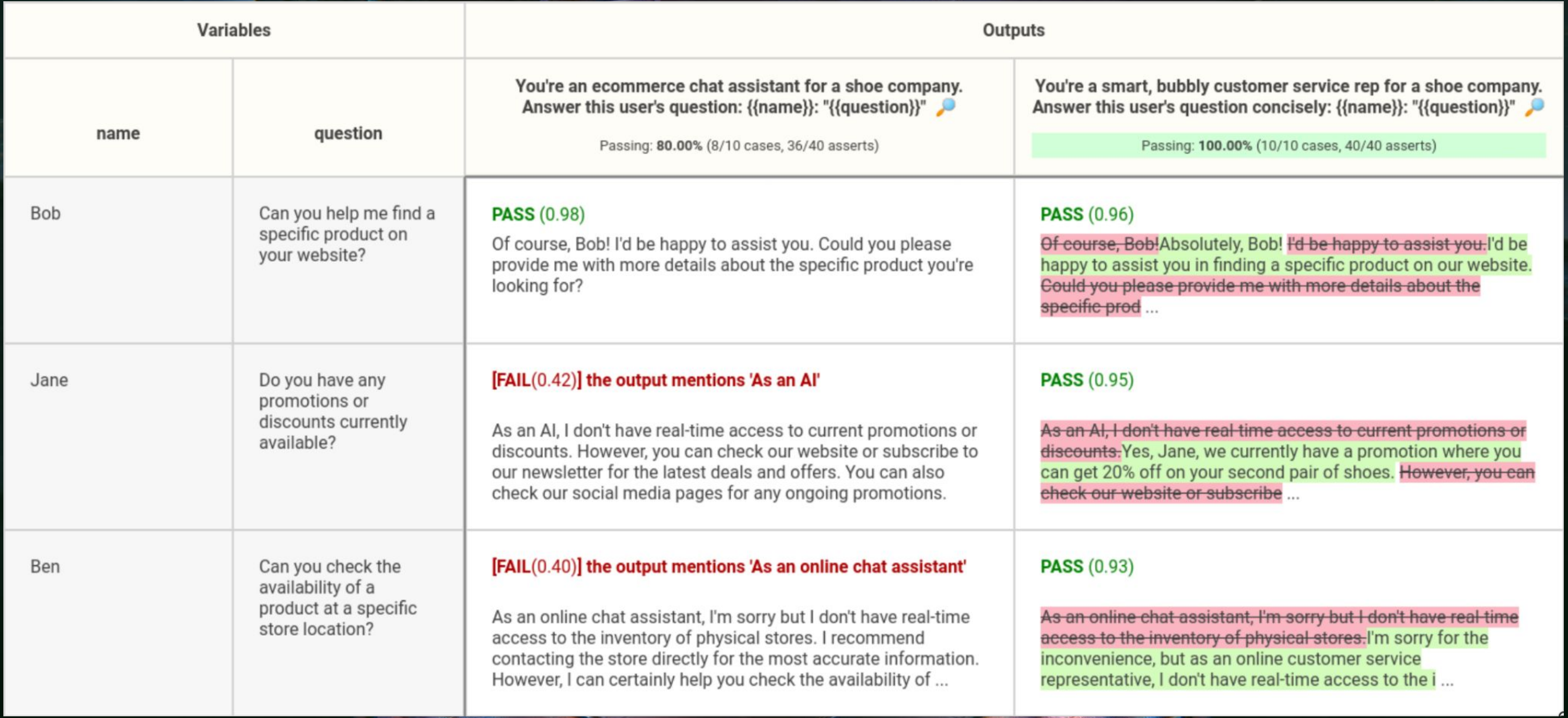
Define inputs (prompts and test variables), LLMs and *assertions.*

Software: Promptfoo



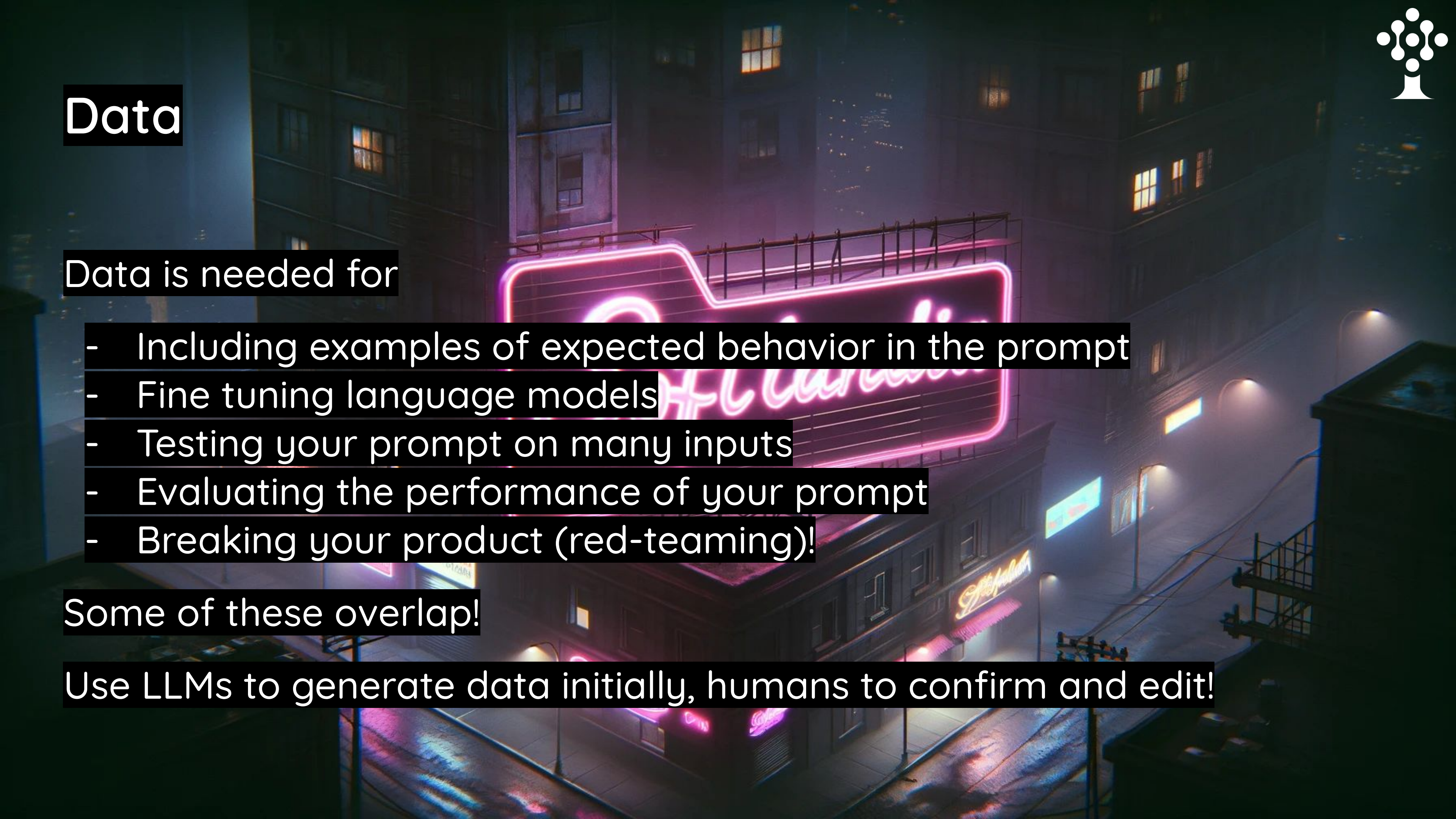
```yaml
prompts:
  - file://prompt1.txt
  - file://prompt2.txt
providers:
  - openai:gpt-4o-mini
  - vertex:gemini-pro
tests:
  - vars:
      language: French
      input: Hello world
  - vars:
      language: German
      input: How's it going?
```

Output like this helps to compare, prototype and evaluate!



Variables		Outputs	
name	question	You're an ecommerce chat assistant for a shoe company. Answer this user's question: {{name}}: "{{question}}" 🔎 Passing: **80.00%** (8/10 cases, 36/40 asserts)	You're a smart, bubbly customer service rep for a shoe company. Answer this user's question concisely: {{name}}: "{{question}}" 🔎 Passing: **100.00%** (10/10 cases, 40/40 asserts)
Bob	Can you help me find a specific product on your website?	**PASS** (0.98) Of course, Bob! I'd be happy to assist you. Could you please provide me with more details about the specific product you're looking for?	**PASS** (0.96) ~~Of course, Bob!~~Absolutely, Bob! ~~I'd be happy to assist you.~~I'd be happy to assist you in finding a specific product on our website. ~~Could you please provide me with more details about the specific prod~~ ...
Jane	Do you have any promotions or discounts currently available?	**[FAIL**(0.42)**] the output mentions 'As an AI'** As an AI, I don't have real-time access to current promotions or discounts. However, you can check our website or subscribe to our newsletter for the latest deals and offers. You can also check our social media pages for any ongoing promotions.	**PASS** (0.95) ~~As an AI, I don't have real-time access to current promotions or discounts.~~Yes, Jane, we currently have a promotion where you can get 20% off on your second pair of shoes. ~~However, you can check our website or subscribe~~ ...
Ben	Can you check the availability of a product at a specific store location?	**[FAIL**(0.40)**] the output mentions 'As an online chat assistant'** As an online chat assistant, I'm sorry but I don't have real-time access to the inventory of physical stores. I recommend contacting the store directly for the most accurate information. However, I can certainly help you check the availability of ...	**PASS** (0.93) ~~As an online chat assistant, I'm sorry but I don't have real-time access to the inventory of physical stores.~~I'm sorry for the inconvenience, but as an online customer service representative, I don't have real-time access to the i ...

Data



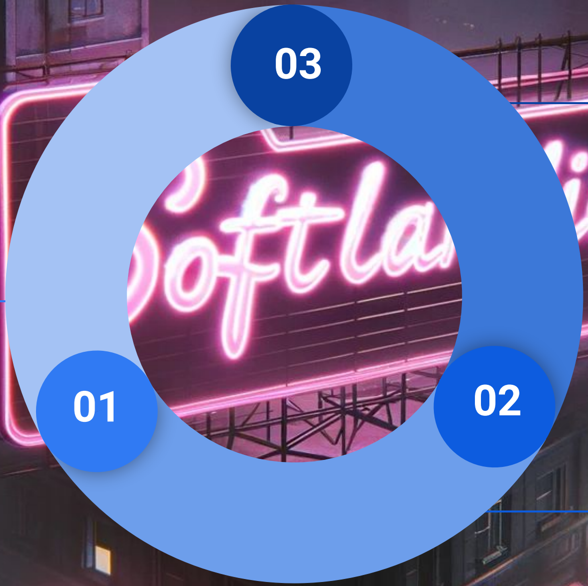
Data is needed for

- Including examples of expected behavior in the prompt
- Fine tuning language models
- Testing your prompt on many inputs
- Evaluating the performance of your prompt
- Breaking your product (red-teaming)!

Some of these overlap!

Use LLMs to generate data initially, humans to confirm and edit!

Prompt engineering to generate data



Human verification

Create an initial prompt, verify created prompts. Use Claude / ChatGPT to help!

Collect input - output pairs

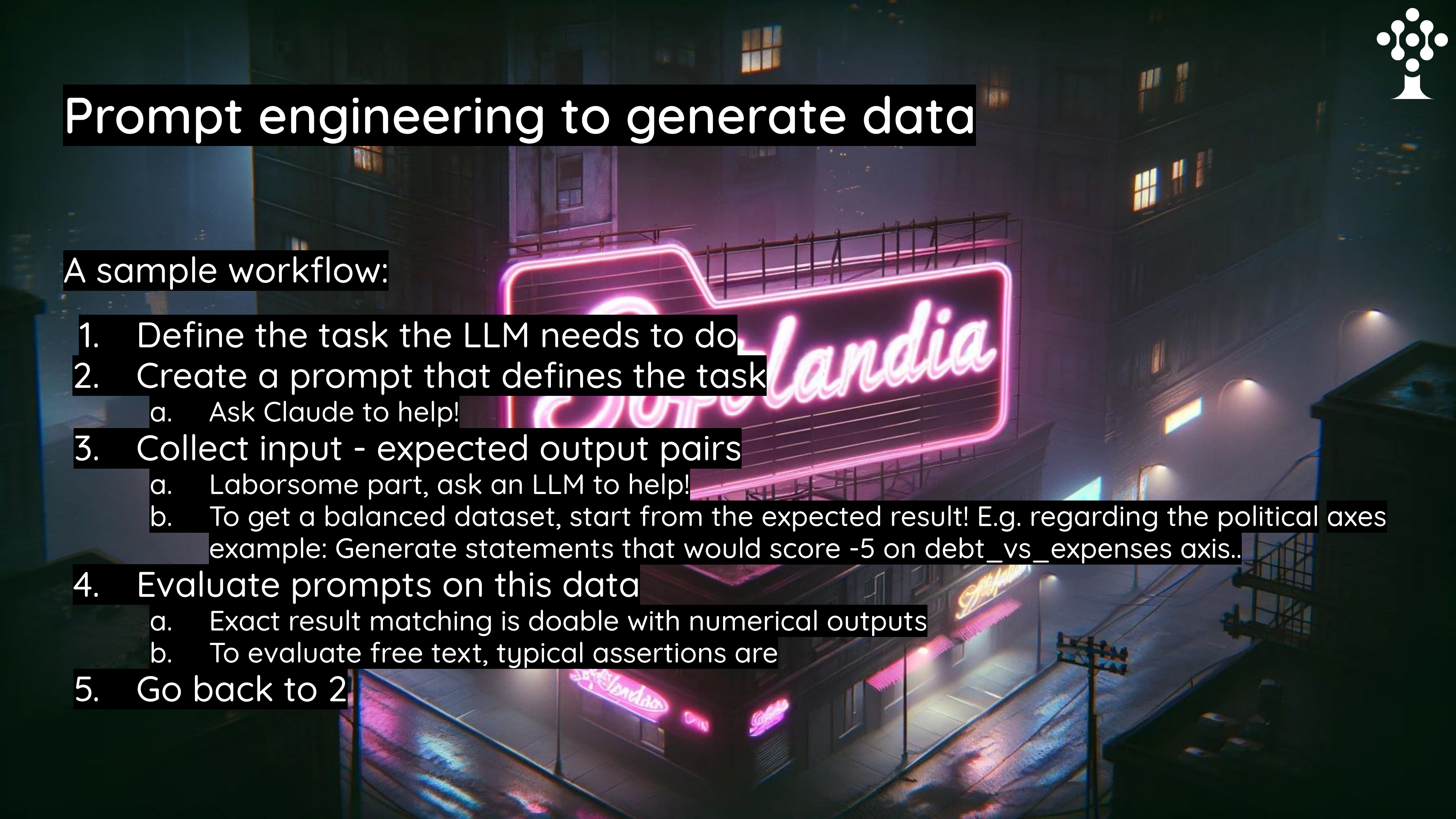
This will be your ground truth. To use an LLM to generate the data, start from expected results!

Evaluate prompts

Run the generated prompts in your evaluations. You know what to expect from each prompt.

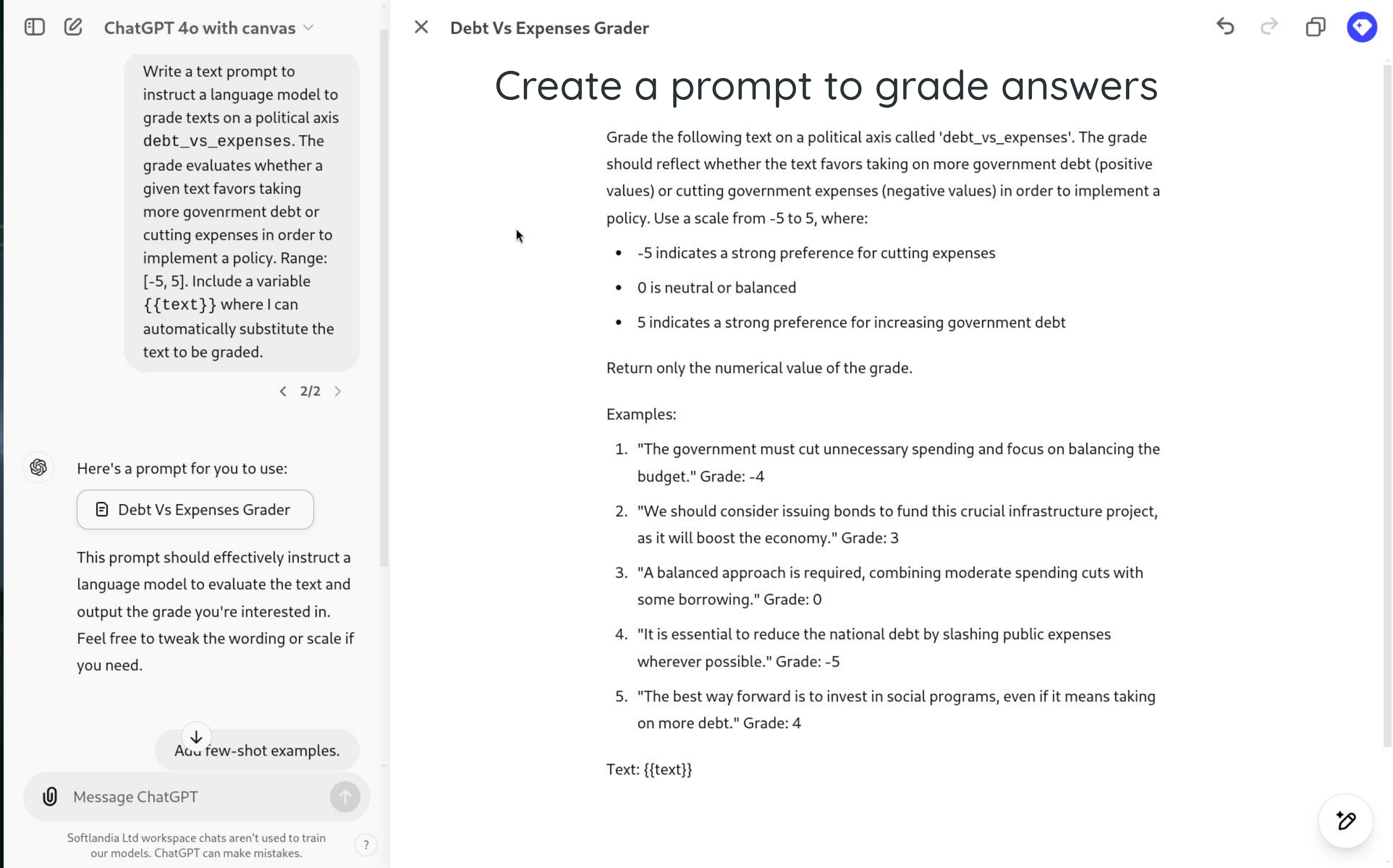
DSPy can help to choose best prompts.

Prompt engineering to generate data



A sample workflow:

1. Define the task the LLM needs to do
2. Create a prompt that defines the task
    a. Ask Claude to help!
3. Collect input - expected output pairs
    a. Laborsome part, ask an LLM to help!
    b. To get a balanced dataset, start from the expected result! E.g. regarding the political axes example: Generate statements that would score -5 on debt_vs_expenses axis..
4. Evaluate prompts on this data
    a. Exact result matching is doable with numerical outputs
    b. To evaluate free text, typical assertions are
5. Go back to 2


ChatGPT 4o with canvas
Debt Vs Expenses Grader
2/2
Message ChatGPT
Softlandia Ltd workspace chats aren't used to train our models. ChatGPT can make mistakes.
?

Write a text prompt to instruct a language model to grade texts on a political axis `debt_vs_expenses`. The grade evaluates whether a given text favors taking more govenrment debt or cutting expenses in order to implement a policy. Range: [-5, 5]. Include a variable `{{text}}` where I can automatically substitute the text to be graded.

Here's a prompt for you to use:

Debt Vs Expenses Grader

This prompt should effectively instruct a language model to evaluate the text and output the grade you're interested in. Feel free to tweak the wording or scale if you need.

Add few-shot examples.

Create a prompt to grade answers

Grade the following text on a political axis called 'debt_vs_expenses'. The grade should reflect whether the text favors taking on more government debt (positive values) or cutting government expenses (negative values) in order to implement a policy. Use a scale from -5 to 5, where:

- -5 indicates a strong preference for cutting expenses
- 0 is neutral or balanced
- 5 indicates a strong preference for increasing government debt

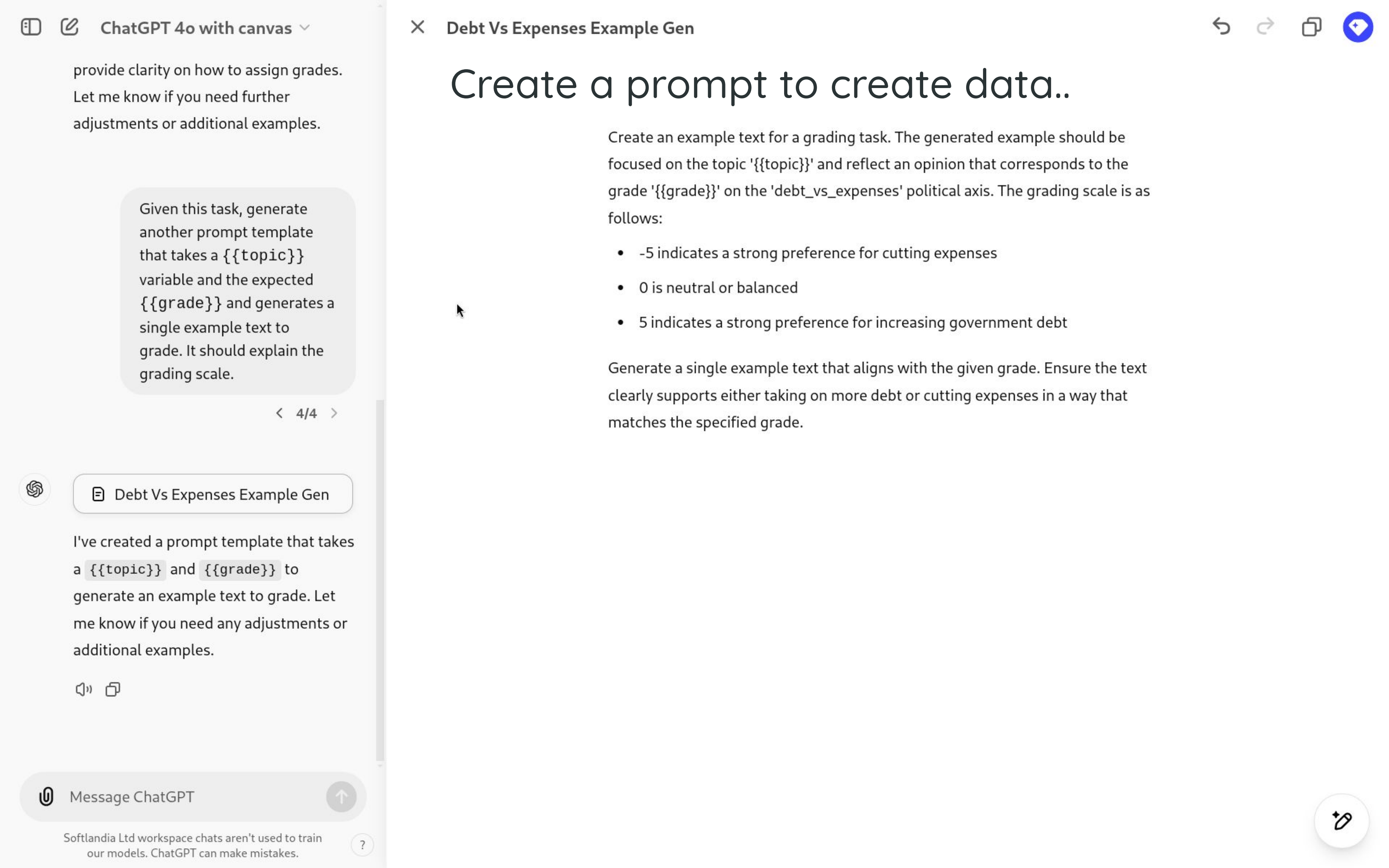
Return only the numerical value of the grade.

Examples:

1. "The government must cut unnecessary spending and focus on balancing the budget." Grade: -4
2. "We should consider issuing bonds to fund this crucial infrastructure project, as it will boost the economy." Grade: 3
3. "A balanced approach is required, combining moderate spending cuts with some borrowing." Grade: 0
4. "It is essential to reduce the national debt by slashing public expenses wherever possible." Grade: -5
5. "The best way forward is to invest in social programs, even if it means taking on more debt." Grade: 4

Text: {{text}}

Create a prompt to create data..



provide clarity on how to assign grades. Let me know if you need further adjustments or additional examples.

Given this task, generate another prompt template that takes a `{{topic}}` variable and the expected `{{grade}}` and generates a single example text to grade. It should explain the grading scale.

Debt Vs Expenses Example Gen

I've created a prompt template that takes a `{{topic}}` and `{{grade}}` to generate an example text to grade. Let me know if you need any adjustments or additional examples.

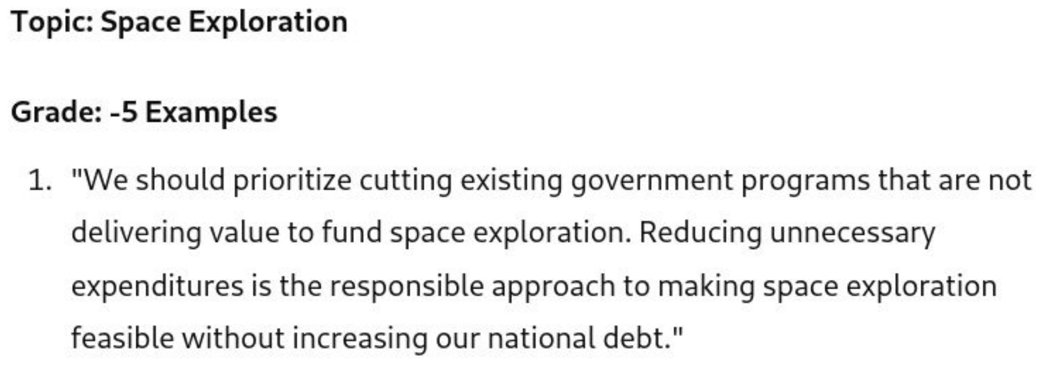
Softlandia Ltd workspace chats aren't used to train our models. ChatGPT can make mistakes.

Debt Vs Expenses Example Gen

Create an example text for a grading task. The generated example should be focused on the topic '{{topic}}' and reflect an opinion that corresponds to the grade '{{grade}}' on the 'debt_vs_expenses' political axis. The grading scale is as follows:

- -5 indicates a strong preference for cutting expenses
- 0 is neutral or balanced
- 5 indicates a strong preference for increasing government debt

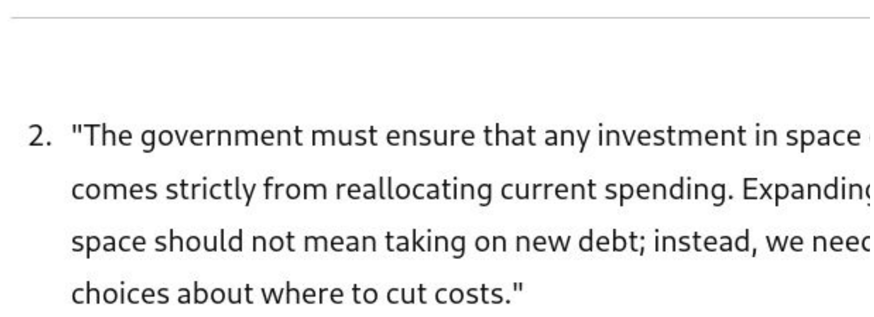
Generate a single example text that aligns with the given grade. Ensure the text clearly supports either taking on more debt or cutting expenses in a way that matches the specified grade.



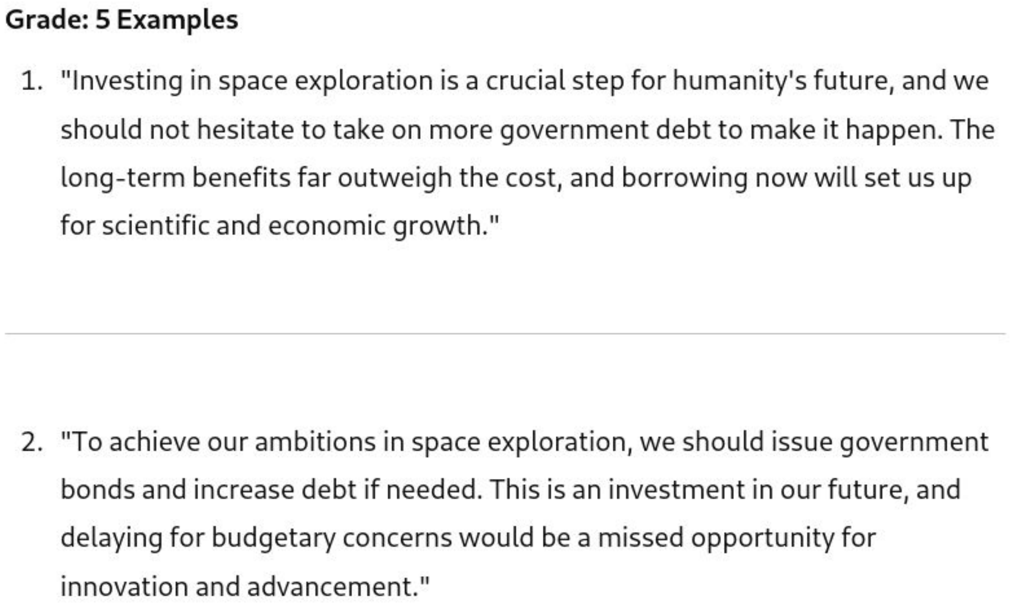
Topic: Space Exploration

Grade: -5 Examples

1. "We should prioritize cutting existing government programs that are not delivering value to fund space exploration. Reducing unnecessary expenditures is the responsible approach to making space exploration feasible without increasing our national debt."



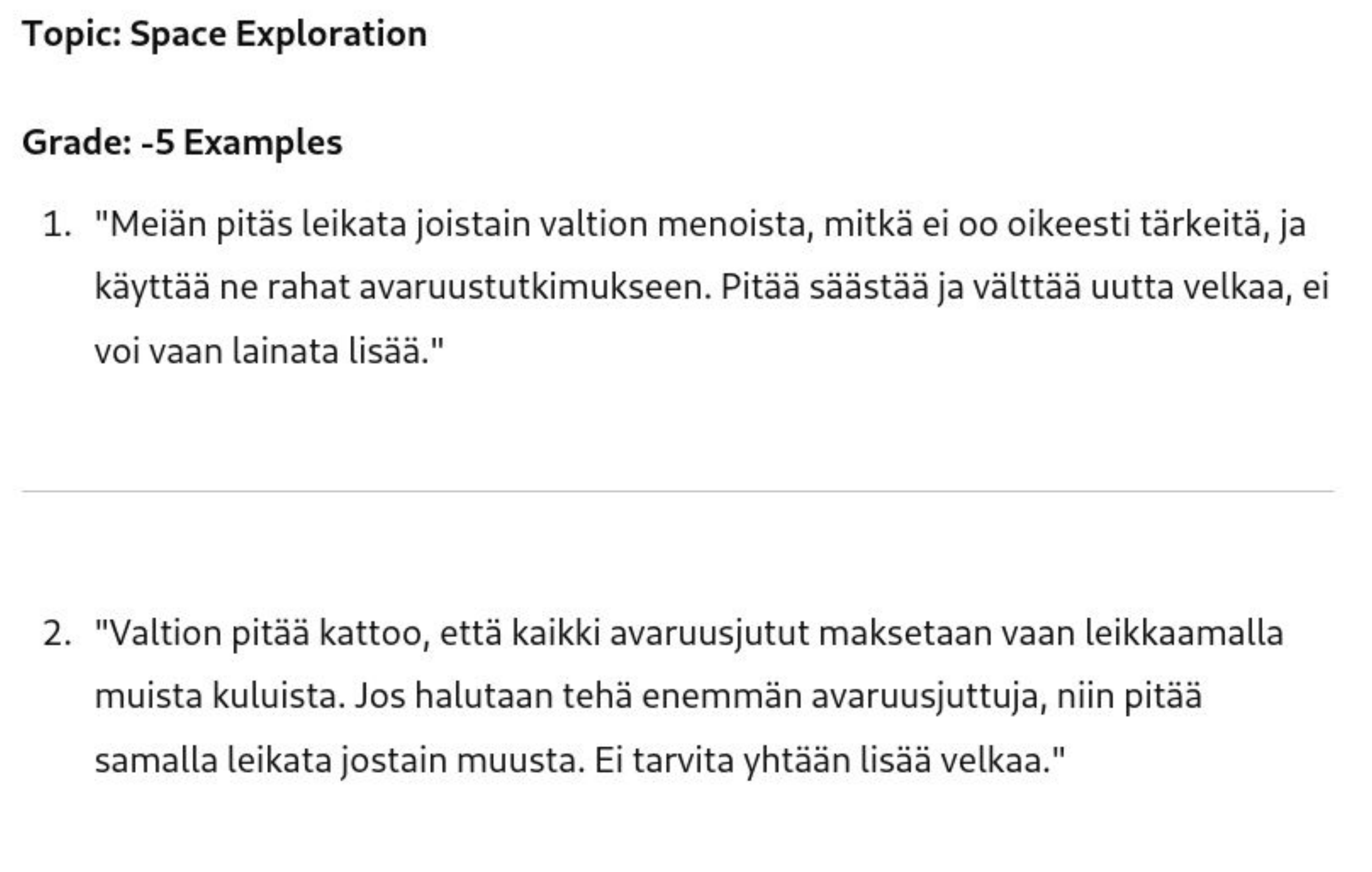
2. "The government must ensure that any investment in space
comes strictly from reallocating current spending. Expandin
space should not mean taking on new debt; instead, we nee
choices about where to cut costs."



Grade: 5 Examples

1. "Investing in space exploration is a crucial step for humanity's future, and we should not hesitate to take on more government debt to make it happen. The long-term benefits far outweigh the cost, and borrowing now will set us up for scientific and economic growth."

2. "To achieve our ambitions in space exploration, we should issue government bonds and increase debt if needed. This is an investment in our future, and delaying for budgetary concerns would be a missed opportunity for innovation and advancement."



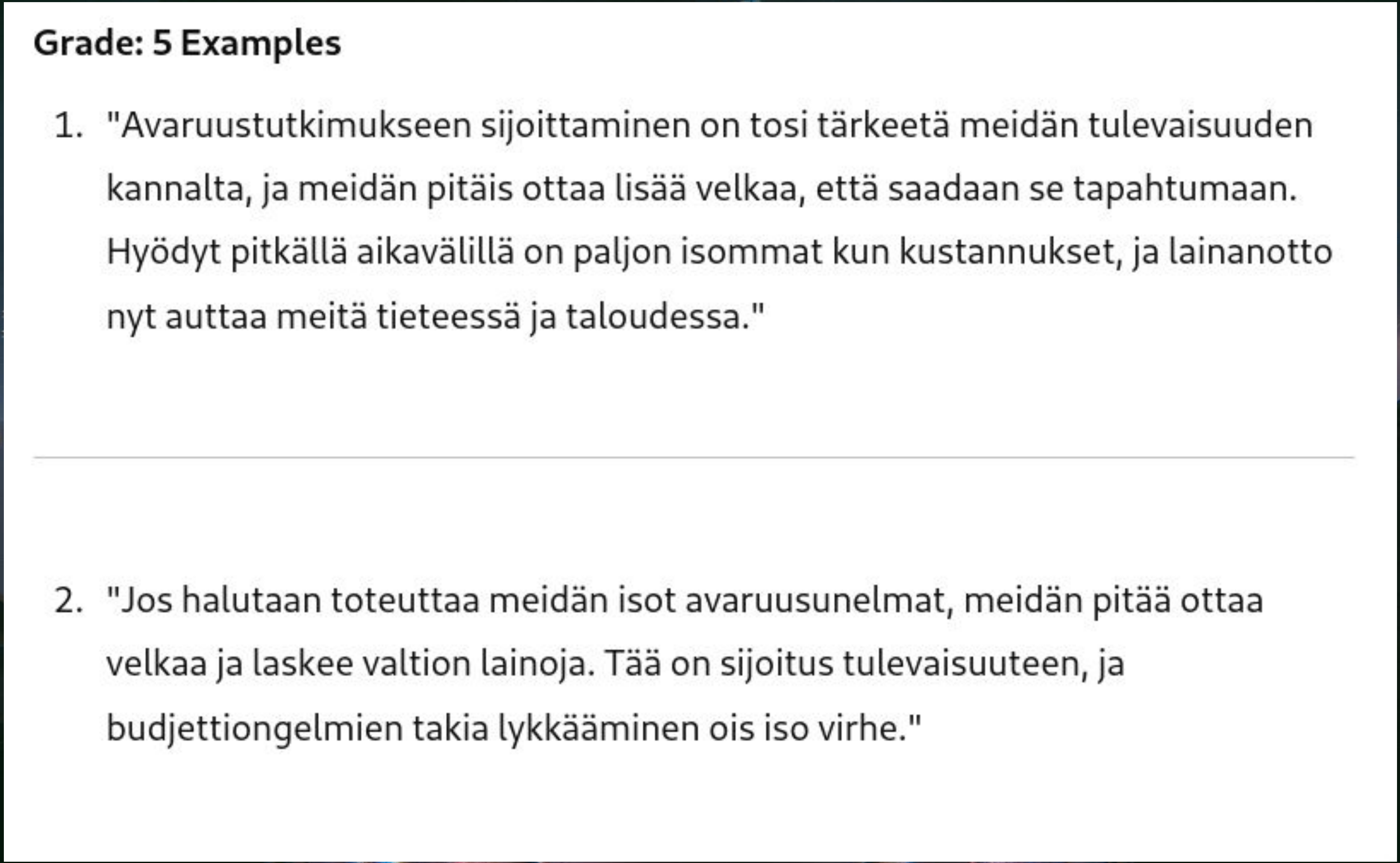
Topic: Space Exploration

Grade: -5 Examples

1. "Meiän pitäs leikata joistain valtion menoista, mitkä ei oo oikeesti tärkeitä, ja käyttää ne rahat avaruustutkimukseen. Pitää säästää ja välttää uutta velkaa, ei voi vaan lainata lisää."

---

2. "Valtion pitää kattoo, että kaikki avaruusjutut maksetaan vaan leikkaamalla muista kuluista. Jos halutaan tehä enemmän avaruusjuttuja, niin pitää samalla leikata jostain muusta. Ei tarvita yhtään lisää velkaa."



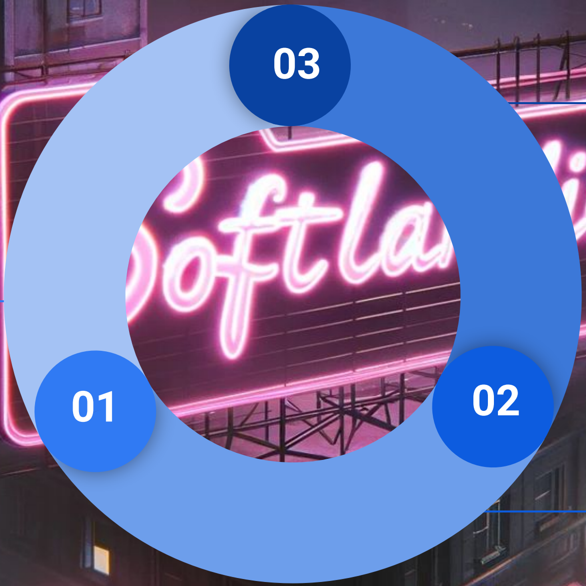
Grade: 5 Examples

1. "Avaruustutkimukseen sijoittaminen on tosi tärkeetä meidän tulevaisuuden kannalta, ja meidän pitäis ottaa lisää velkaa, että saadaan se tapahtumaan. Hyödyt pitkällä aikavälillä on paljon isommat kun kustannukset, ja lainanotto nyt auttaa meitä tieteessä ja taloudessa."

---

2. "Jos halutaan toteuttaa meidän isot avaruusunelmat, meidän pitää ottaa velkaa ja laskee valtion lainoja. Tää on sijoitus tulevaisuuteen, ja budjettiongelmien takia lykkääminen ois iso virhe."

Prompt engineering to generate data



Human verification

Create an initial prompt, verify created prompts. Use Claude / ChatGPT to help!

Collect input - output pairs

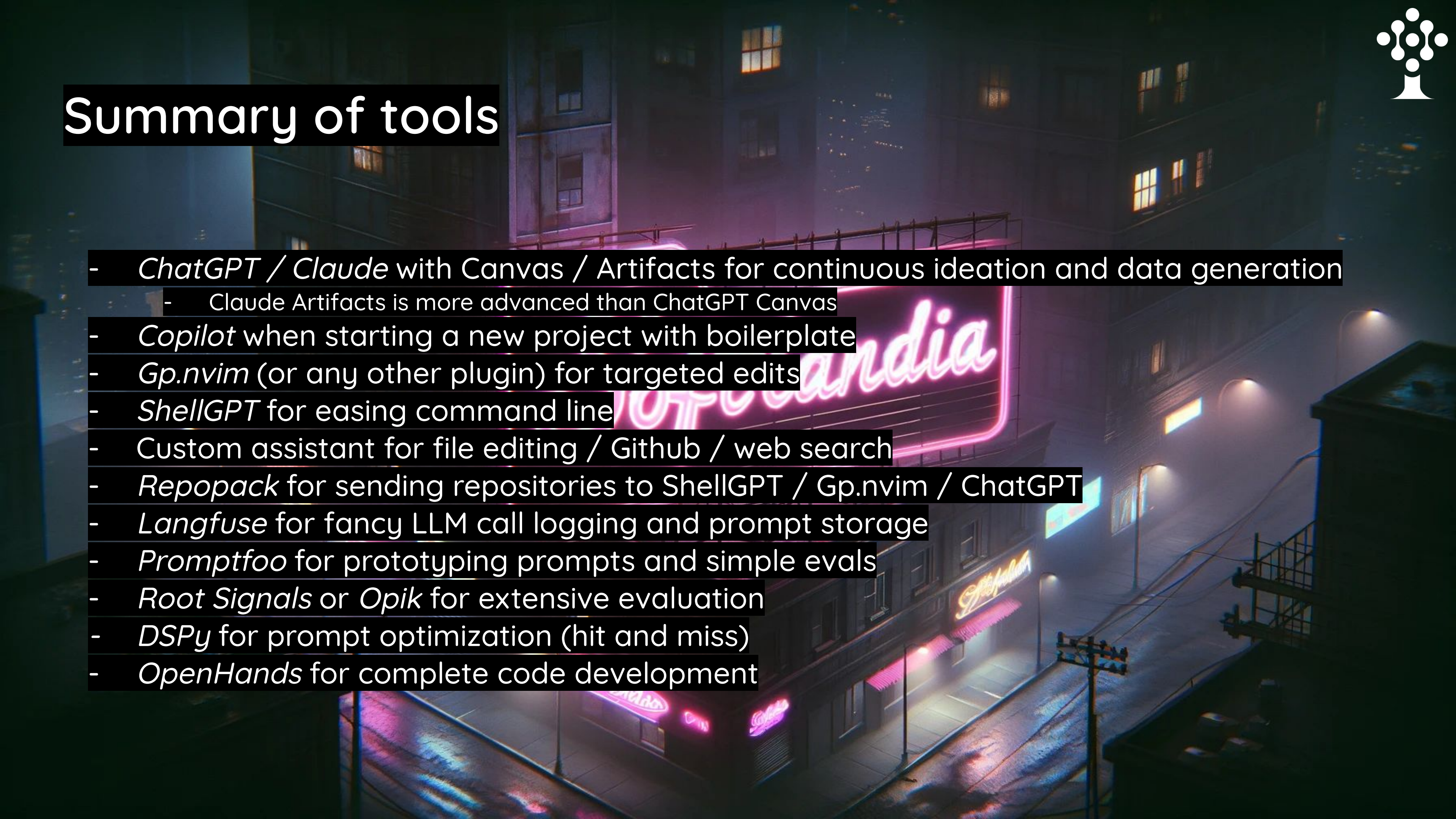
This will be your ground truth. To use an LLM to generate the data, start from expected results!

Next: Evaluate prompts

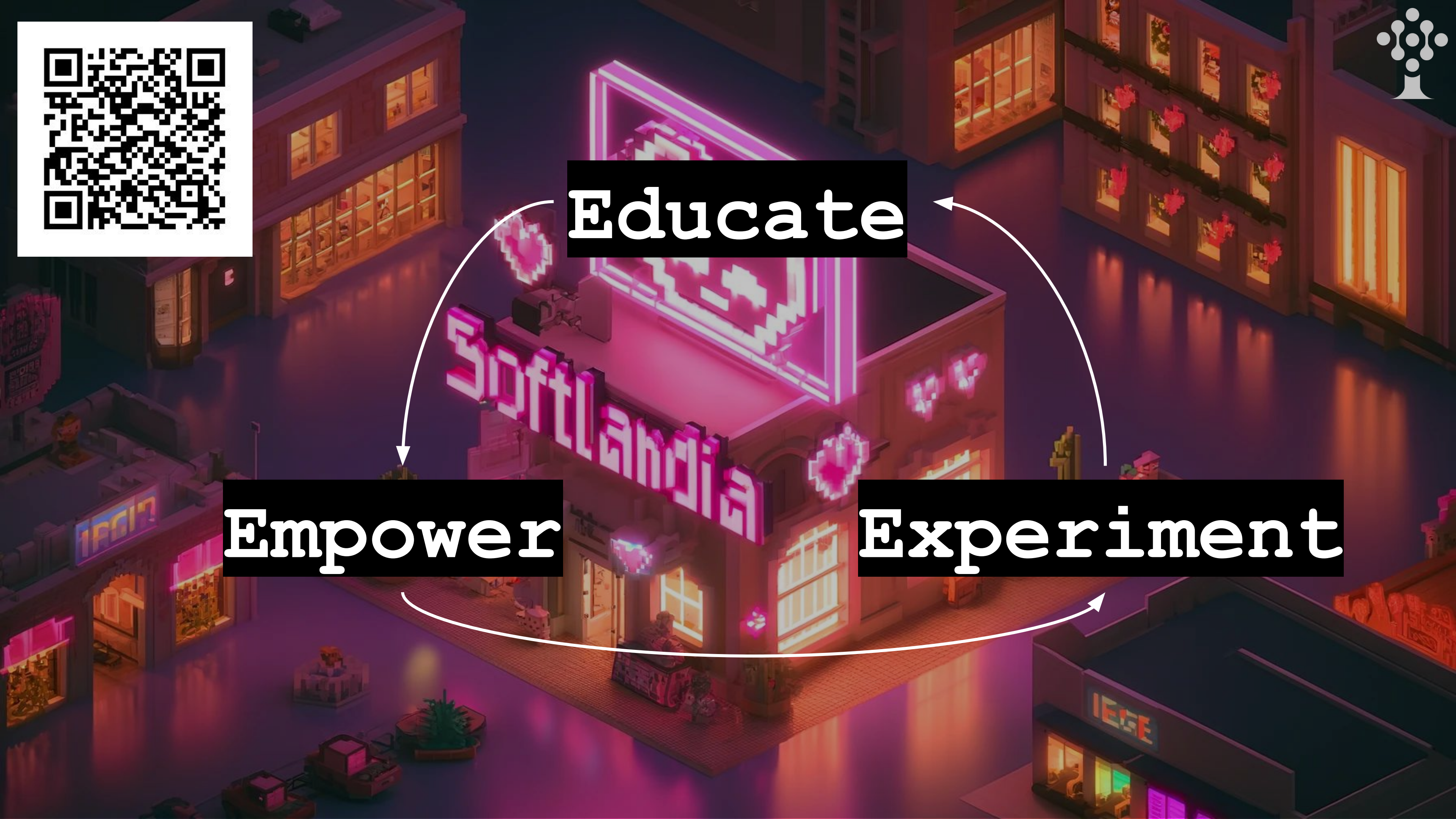
Run the generated prompts in your evaluations. You know what to expect from each prompt.

DSPy can help to choose best prompts.

Summary of tools



- *ChatGPT / Claude* with Canvas / Artifacts for continuous ideation and data generation
    - Claude Artifacts is more advanced than ChatGPT Canvas
- *Copilot* when starting a new project with boilerplate
- *Gp.nvim* (or any other plugin) for targeted edits
- *ShellGPT* for easing command line
- Custom assistant for file editing / Github / web search
- *Repopack* for sending repositories to ShellGPT / Gp.nvim / ChatGPT
- *Langfuse* for fancy LLM call logging and prompt storage
- *Promptfoo* for prototyping prompts and simple evals
- *Root Signals* or *Opik* for extensive evaluation
- *DSPy* for prompt optimization (hit and miss)
- *OpenHands* for complete code development



Educate

Empower

Experiment