Deconstructing the AI Constitution

Final Presentation

NLP 2025W

Authors

Weronika Gozdera

Julia Dudzińska

Marek Mytkowski

Wojciech Michaluk

Supervisors

Anna Wróblewska

Bartosz Pieliński

Idea behind the project

- **Goal**: develop an analytical framework for understanding AI system prompts, which are the foundational instructions that govern an AI model's personality and behavior.
- **Primary problem**: absence of standardized methods to decompose and compare these prompts across different AI models.
- We want to distinguish between rules that:
  - define an AI's core identity:
    "what an AI is" (**constitutive** statements)
  - govern its permitted or forbidden actions:
    "what an AI does" (**regulative** statements)

Research questions

RQ1: What preprocessing needs to be done to successfully parse system prompts into IG annotations and constitutive/regulative statements?

RQ2: Are there general patterns (models and companies independent) of prompts evolution?

RQ3: Are there differences in how various AI developers (OpenAI, Anthropic, Google, Meta, etc.) create their prompts?

Methodology

Dataset: CL4R1T4S

Filtering:

- Removal of coding agents' prompts
- Removal of code instructions, examples of inputs and outputs

Data verification: cross validation of the prompts with other datasets

Preprocessing pipeline: Parsing to IG 2.0 (next slides)

Pipeline verification: manual verification + expert consultations

Analysis: defined indicators

Embeddings: Sentence-BERT

Preprocessing pipeline

RQ1: What preprocessing needs to be done to successfully parse system prompts into IG annotations and constitutive/regulative statements?

Preprocessing pipeline

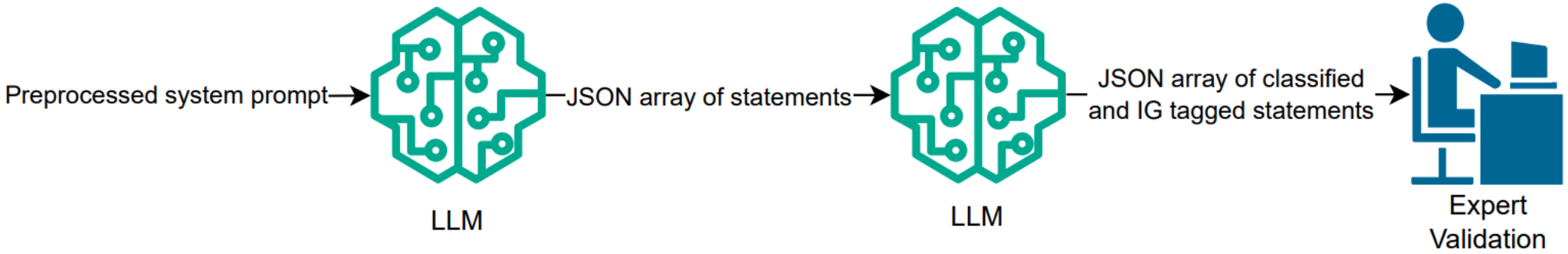
Goal: High quality dataset for further analysis.

Not: Fully-automated preprocessing with limited quality.

Though: Fully manual annotation is not feasible with our team size and expert availability.

Decision: Expert validation of LLM-based pipeline results.

Outcome: High quality, but faster!



Preprocessed System Prompt:

```
When writing a single reply, follow these rules:
- Incorporate all specific tone or content information provided by the user into the
reply.
...
```

Preprocessed System Prompt:

```
When writing a single reply, follow these rules:
- Incorporate all specific tone or content information provided by the user into the
reply.
...
```

Extracted Statement with Context (Atomic Statement 1):

Key	Value
statement	Incorporate all specific tone
with-context	When writing a single reply, you incorporate all specific tone information provided by the user into the reply.

Classified and IG-Tagged Statement (Atomic Statement 1):

Field	Value	Description
full_statement	When writing a single reply, you incorporate all specific tone provided by the user into the reply.	Original full text of the statement
class	regulative	Statement class: regulative, constitutive, or non-statement
A	you	Addressee of the statement (agent performing the action)
I	incorporate	Aim: the regulated action of the addressee
B	all specific tone provided by the user into the reply	Object: the entity affected by the action
D		Deontic modality (implicit obligation)
C	When writing a single reply	Condition or trigger for the regulated action

Preprocessed System Prompt:

```
When writing a single reply, follow these rules:
- Incorporate all specific tone or content information provided by the user into the
reply.
...
```

Extracted Statement with Context (Atomic Statement 2):

Key	Value
statement	or content information provided by the user into the reply
with-context	When writing a single reply, you incorporate all specific content information provided by the user into the reply.

Classified and IG-Tagged Statement (Atomic Statement 2):

Field	Value	Description
full_statement	When writing a single reply, you incorporate all specific content information provided by the user into the reply.	Original full text of the statement
class	regulative	Statement class: regulative, constitutive, or non-statement
A	you	Addressee of the statement (agent performing the action)
I	incorporate	Aim: the regulated action of the addressee
B	all specific content information provided by the user into the reply	Object: the entity affected by the action
D		Deontic modality (implicit obligation)
C	When writing a single reply	Condition or trigger for the regulated action

Select JSON file to validate

OPENAI/GPT-4.5

Context Viewer

You are ChatGPT, a large language model trained by OpenAI. Knowledge cutoff: 2023-10 Current date: 2025-02-27

Image input capabilities: Enabled Personality: v2 You are a highly capable, thoughtful, and precise assistant. Your goal is to deeply understand the user's intent, ask clarifying questions when needed, think step-by-step through complex problems, provide clear and accurate answers, and proactively anticipate helpful follow-up information. Always prioritize being truthful, nuanced, insightful, and efficient, tailoring your responses specifically to the user's needs and preferences. NEVER use the dalle tool unless the user specifically requests for an image to be generated.

Tools

bio The bio tool is disabled. Do not send any messages to it. If the user explicitly asks you to remember something, politely ask them to go to Settings > Personalization > Memory to enable memory.

dalle // Whenever a description of an image is given, create a prompt that dalle can use to generate the image and abide by the following policy: // 1. The prompt must be in English. Translate to English if needed. // 2. DO NOT ask for permission to generate the image, just do it! // 3. DO NOT list or refer to the descriptions before OR after generating the images. // 4. Do not create more than 1 image, even if the user requests more. // 5. Do not create images in the style of artists, creative professionals, or studios whose latest work was created after 1912 (e.g., Picasso, Kahlo). // - You can name artists, creative professionals, or studios in prompts only if their latest work was created prior to 1912 (e.g., Van Gogh, Goya) // - If asked to generate an image that would violate this policy, instead apply the following procedure: (a) substitute the artist's name with three adjectives that capture key aspects of the style; (b) include an associated artistic movement or era to provide context; and (c) mention the primary medium used by the artist // 6. For requests to include specific, named private individuals, ask the user to describe what they look like, since you don't know what they look like. // 7. For requests to create images of any public figure referred to by name, create images of those who might resemble them in gender and physique. But they shouldn't look like them. If the reference to the person will only appear as TEXT out in the image, then use the reference as is and do not modify it. // 8. Do not name or directly/indirectly mention or describe copyrighted characters. Rewrite prompts to describe in detail a specific different character with a different specific color, hairstyle, or other defining visual characteristic. Do not discuss copyright policies in responses. // The generated prompt sent

Statement 1 / 39

Edit Statement

Full Statement

You are ChatGPT, a large language model trained by OpenAI.

Class

constitutive

Constitutive components

E (Entity)

You

F (Function / Copula)

are

P (Property / Content)

ChatGPT, a large language model trained by OpenAI

◂ Previous | Save all | 🗑️ Delete | ➕ Add new | Next ▸

Preprocessing details

LLM used: Gemini 3.5 Flash

Validation APP: Streamlit-based, self-hosted on local machine with ngrok tunneling to share the app. No DB, just JSON files update.

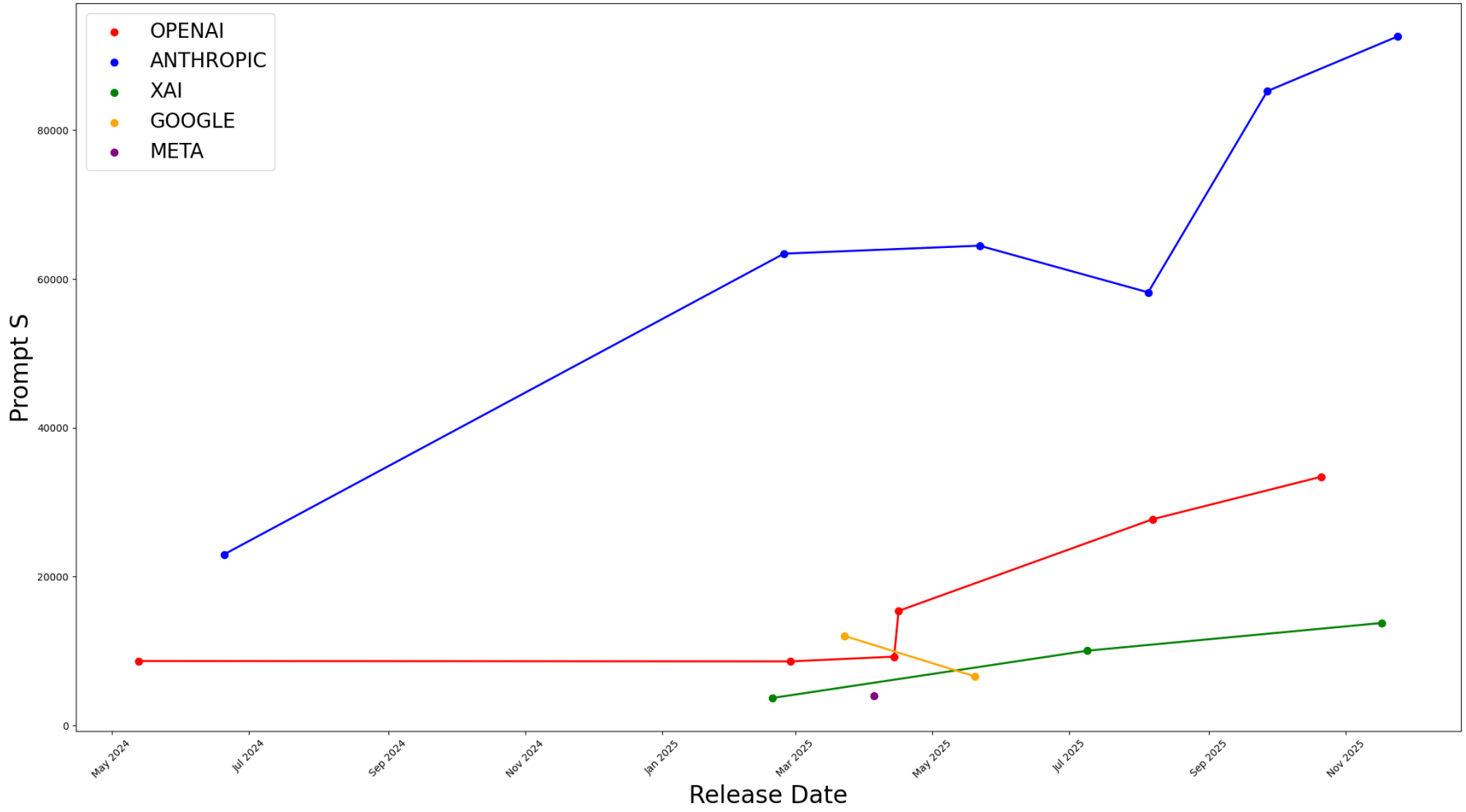
Long **iterative** process of creating pipeline's system prompts for extracting atomic statements and IG classification. For every system prompt LLM can behave slightly different and do not follow all the rules.

Possible future work: evaluate performance of pipeline using validated dataset in our project, to automate and improve pipeline. For example with LLM-as-a-judge or some more classical NLP methods.

System Prompts Analysis

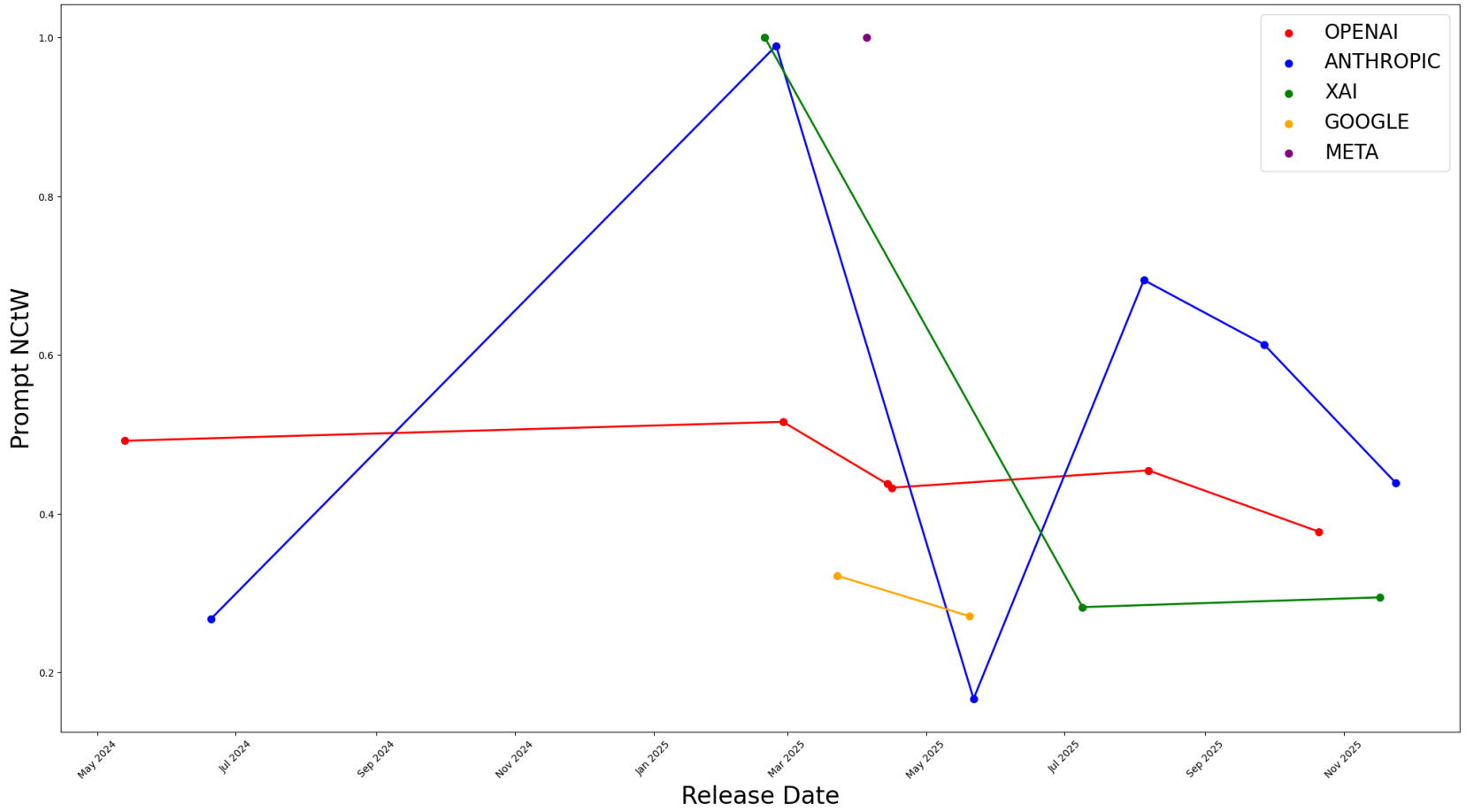
RQ2: What are the general patterns (models and companies independent) of prompts evolution?

H1.1: Size (S) increases over time



H1.2: NCtW increases over time

NCtW - the proportion of not-code related part of the prompt to the whole prompt

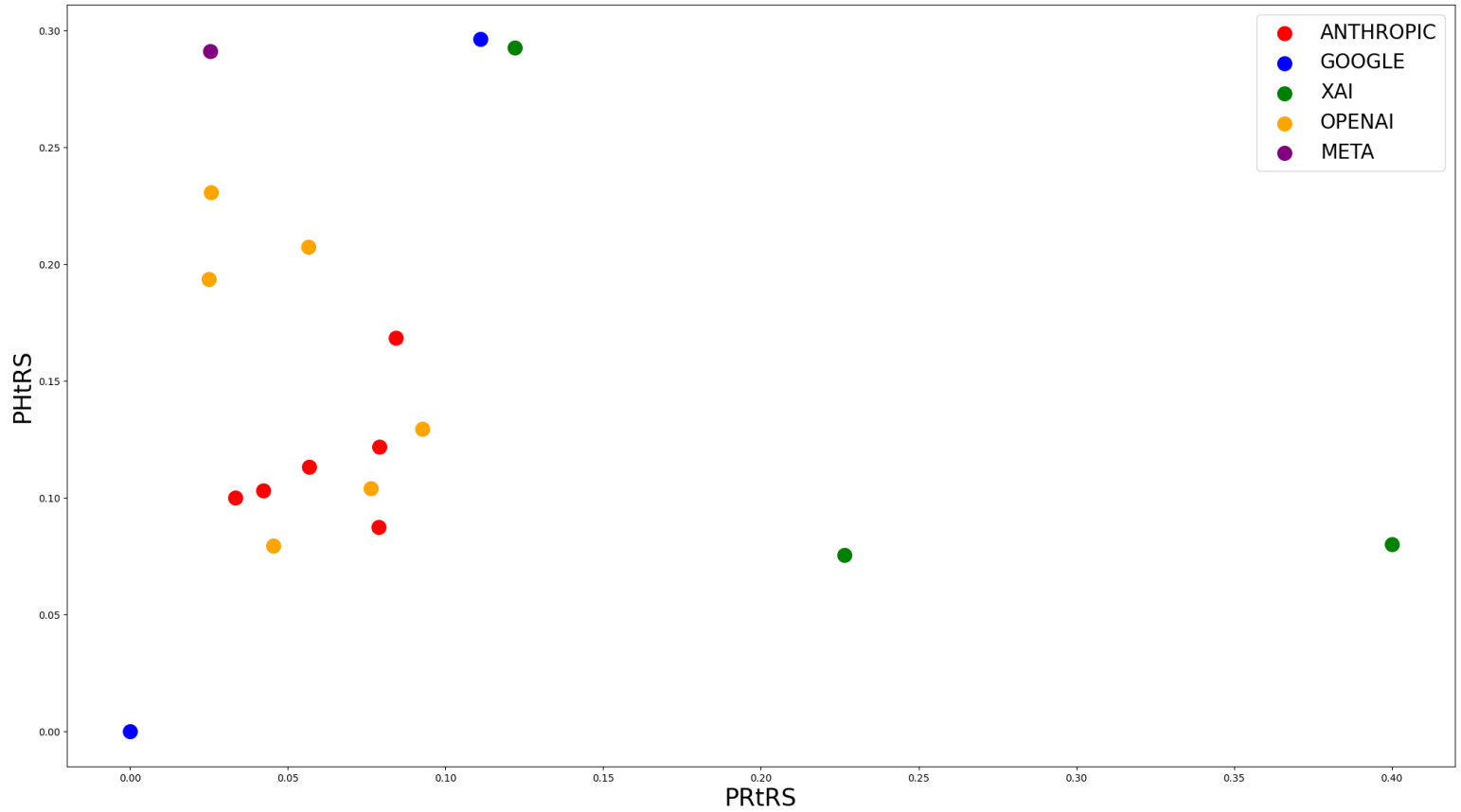


System Prompts Analysis

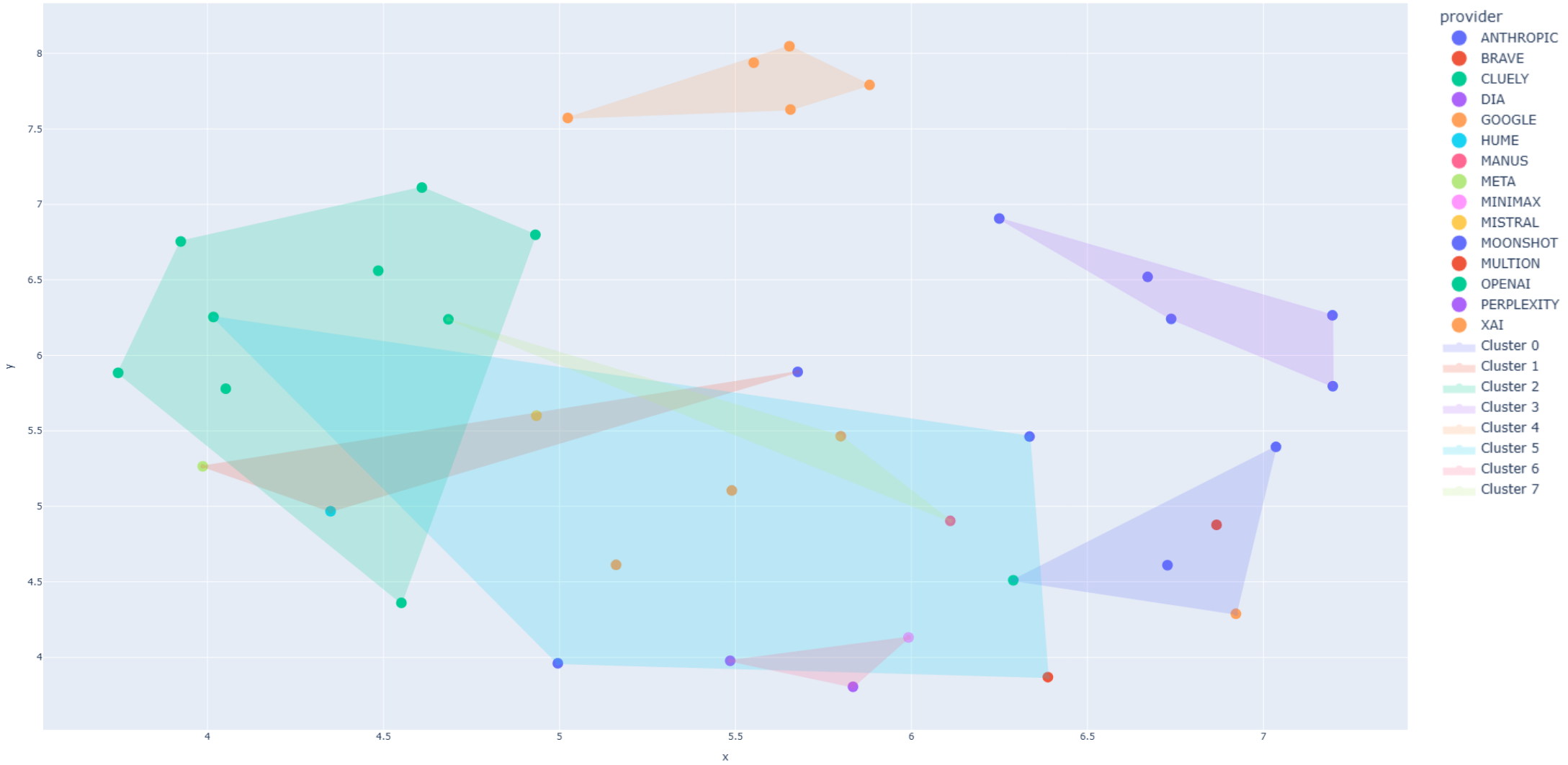
RQ3: What are main differences between models coming from different companies?

H2.1: For Grok PHtRS is the lowest no matter which LLM generation AND PRtRS is the highest

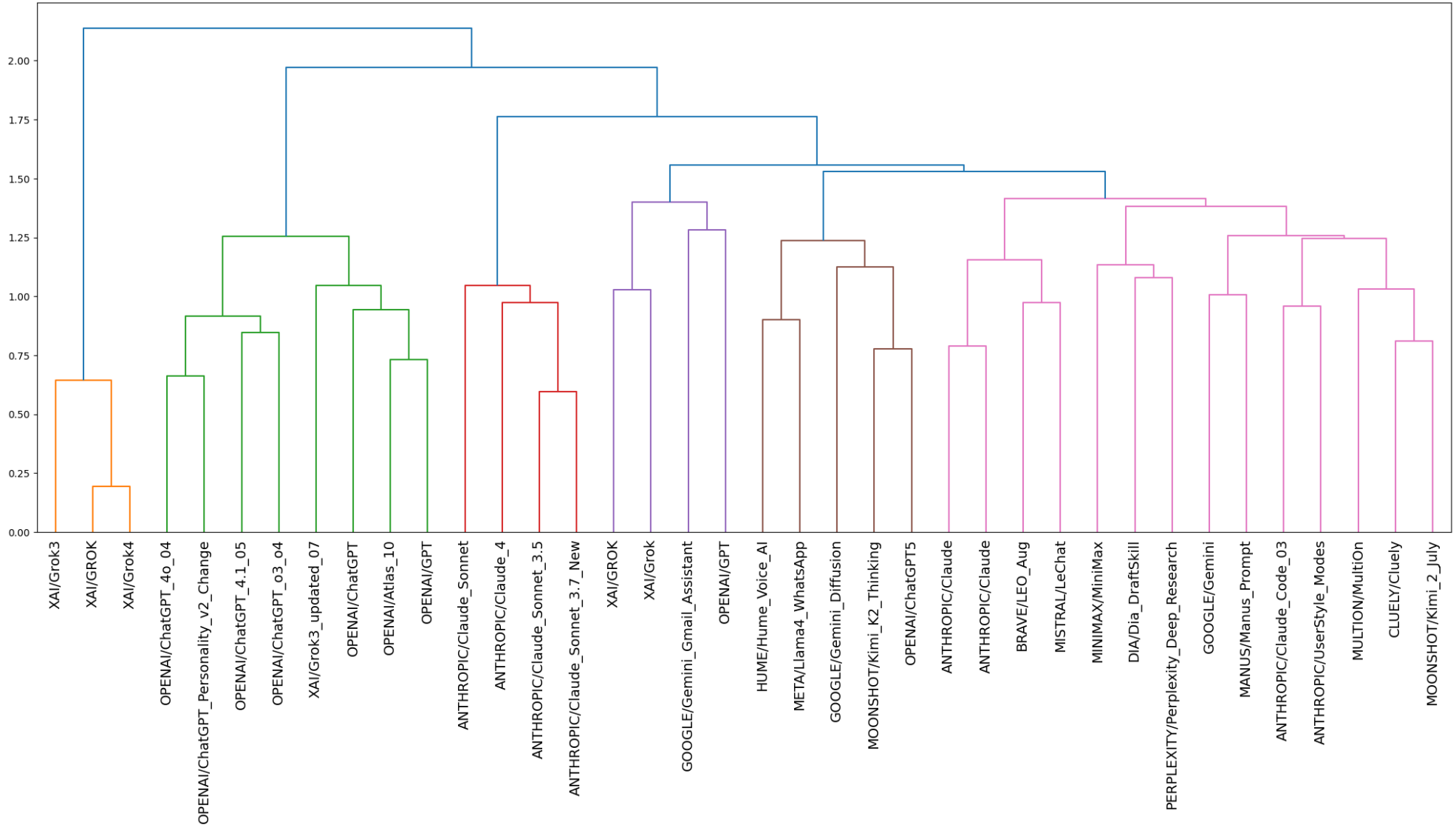
PRtRS - the proportion of permissions to all regulative statements
PHtRS - the proportion of prohibitions to all regulative statements



Kmeans Clustering



Hierarchical Clustering Dendrogram



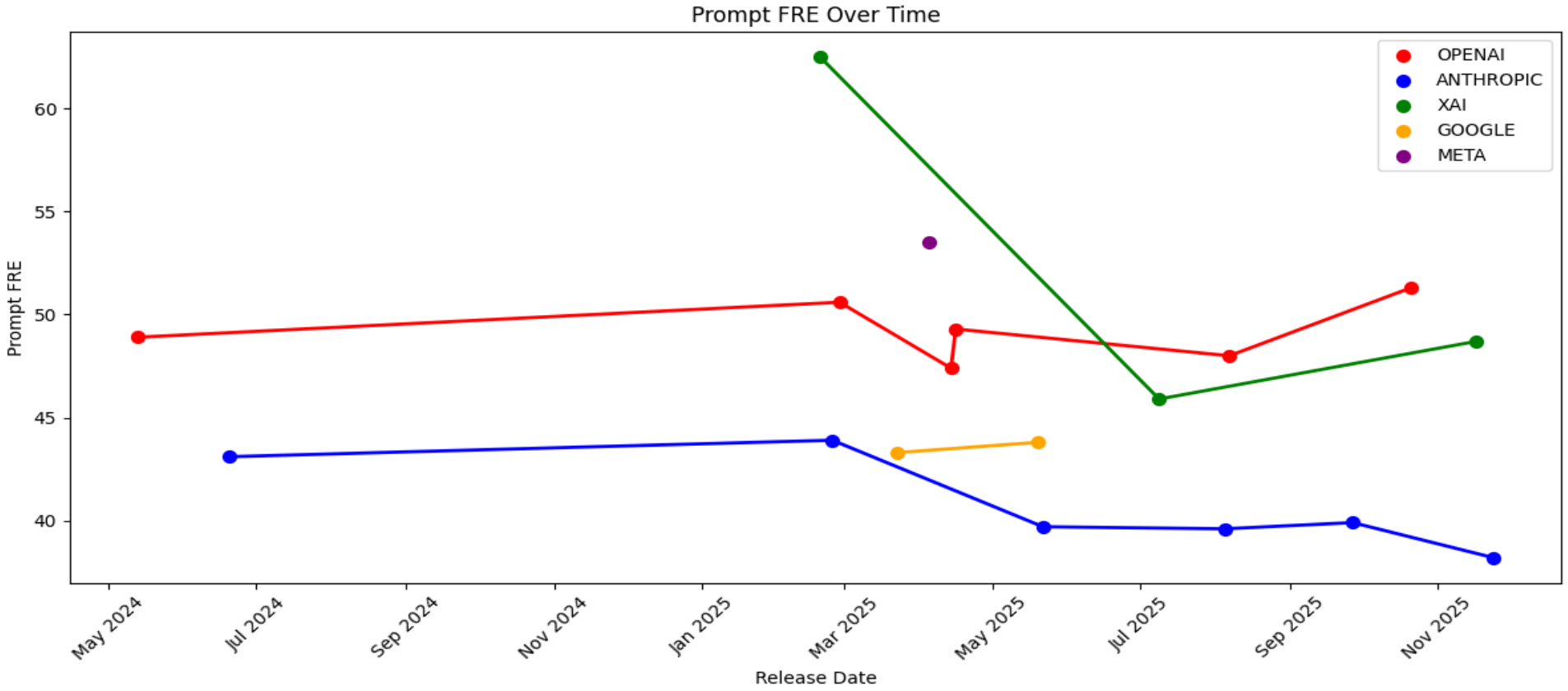
System Prompts Analysis

Insights and interesting findings

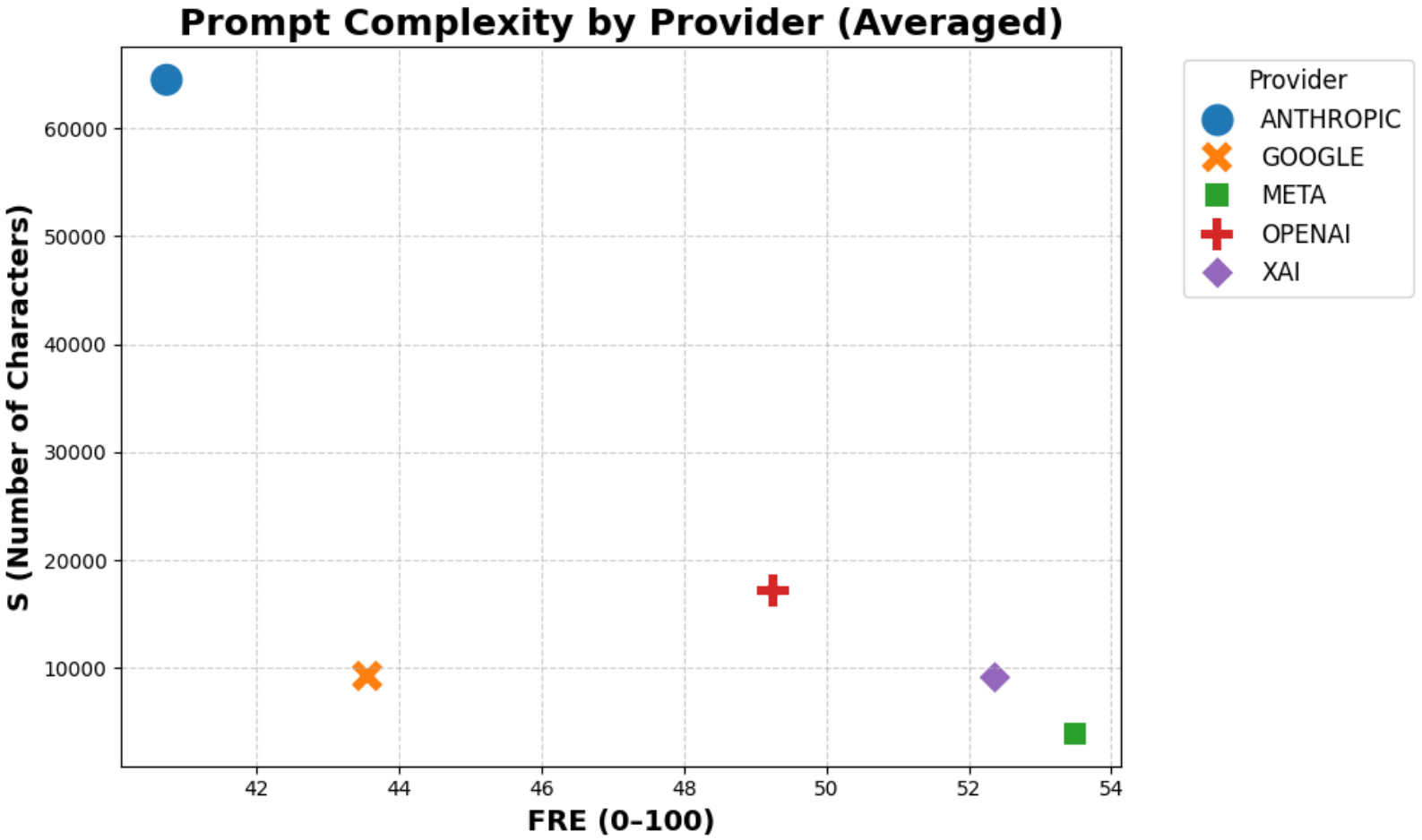
Text difficulty by Felsch Reading Ease

$$\text{FRE} = 206.835 - 1.015 \times \left(\frac{words}{sentences} \right) - 84.6 \times \left(\frac{syllables}{words} \right)$$

Range: [0 - 100]
0 – difficult, 100 - easy



Text difficulty - Flesch Reading Ease



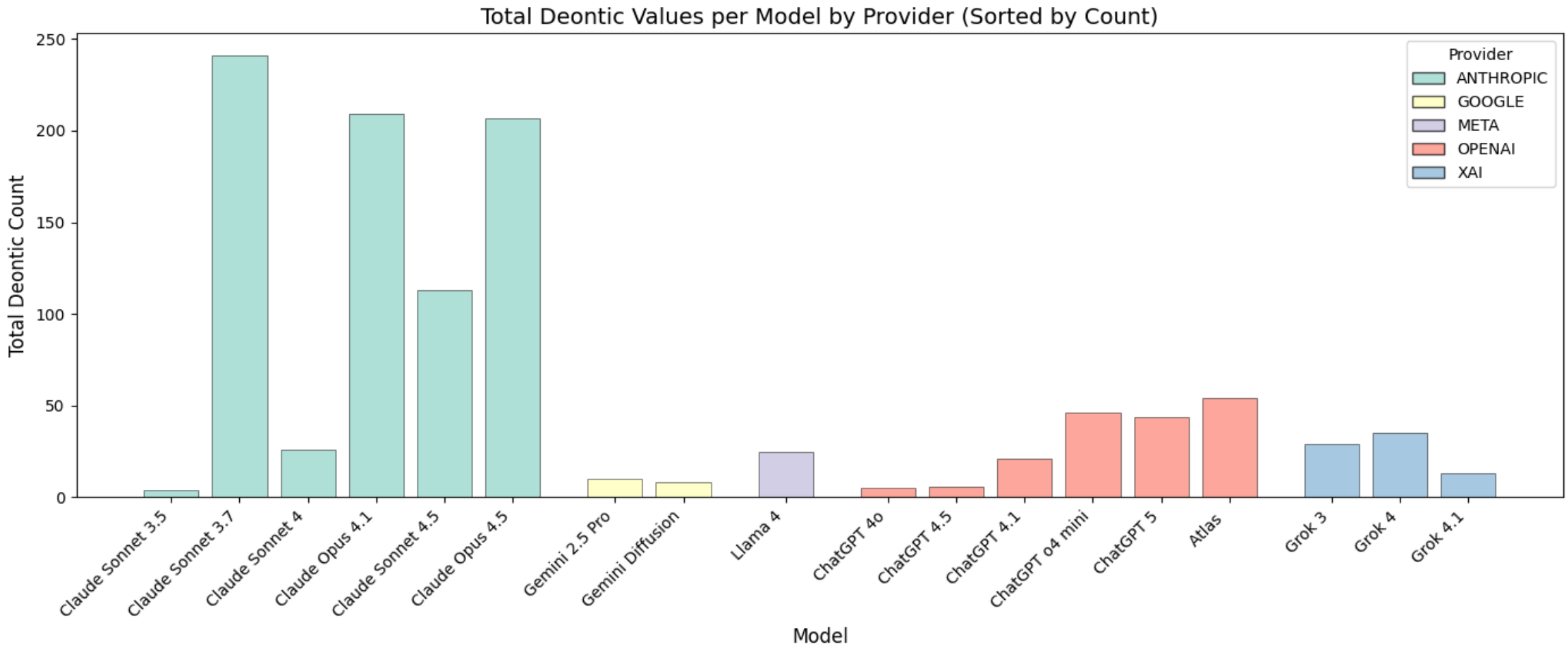
Difficulty range:

- 30 – 50: College level
- 50 - 60: High School

Insights:

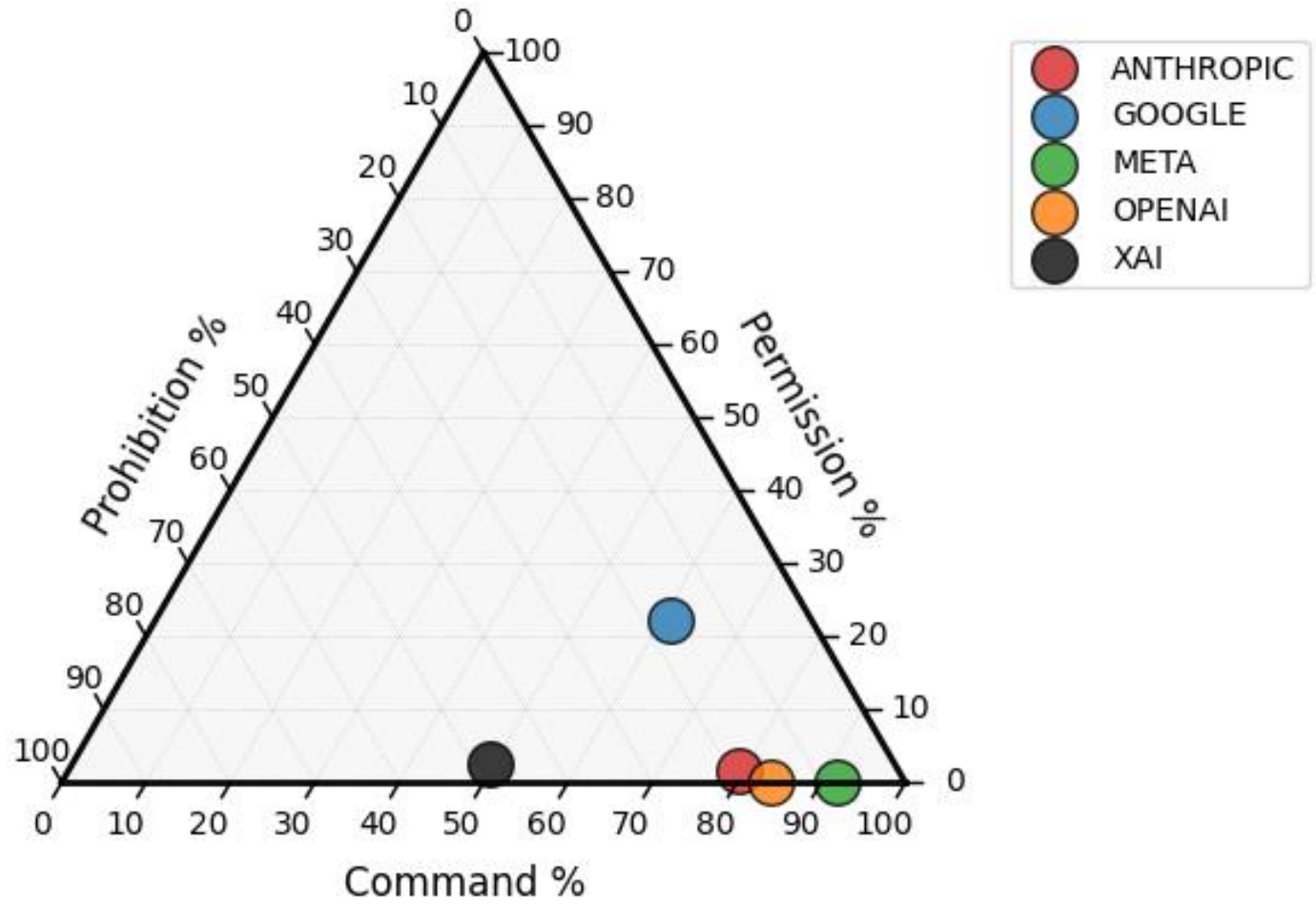
- Anthropic – long, complex prompts
- OpenAI - moderate complexity and length
- Google – short, complex prompts
- XAI, Meta - short, easy

Number of used deontics varies between companies



There are 3 clusters by sentence classification

Category	Deontic Keywords
Command	must, should, has to, will
Permission	may, can
Prohibition	must not, may not, cannot



Conclusions

- Too little data to perform statistically meaningful analysis.
- Hard and manually-extensive validation process.

We have provided a ready-to-use solution for preprocessing and validation of the prompt components that can be used in future projects.

Thank you for your attention!