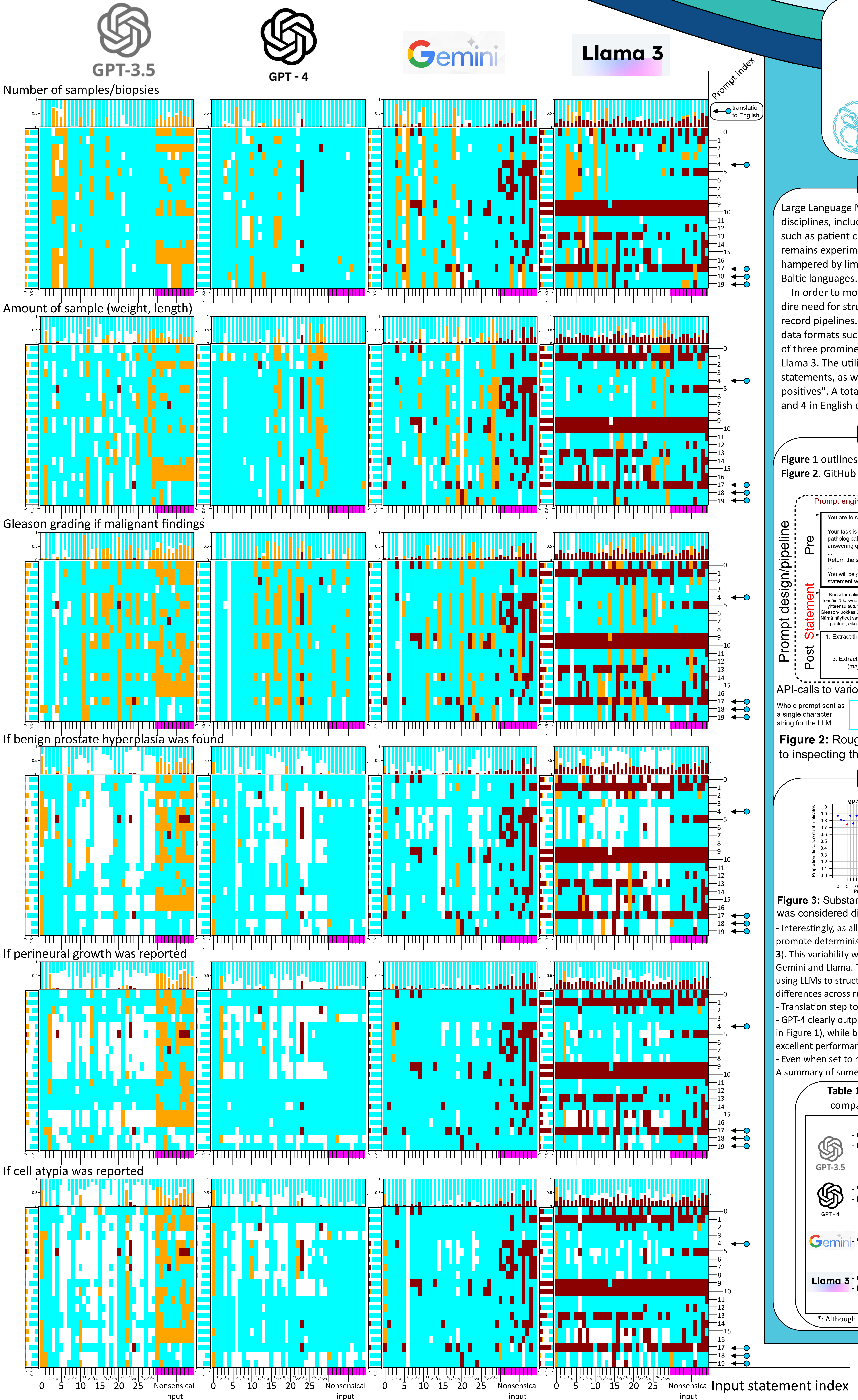
P6

Practical comparison of GPT, Gemini and Llama LLMs for structuring biomedical data from synthetic Finnish prostate cancer pathology statements

Kaleva I.M.[1], Mirtti T.[2], Rannikko A.S.[2], Laajala T.D.[1,2]

1: University of Turku, Department of Mathematics and Statistics, Turku, Finland
2: University of Helsinki, Faculty of Medicine, Helsinki, Finland



Number of samples/biopsies

Amount of sample (weight, length)

Gleason grading if malignant findings

If benign prostate hyperplasia was found

If perineural growth was reported

If cell atypia was reported

Figure 1: 30 synthetic Finnnish histo-pathological statements along with 10 "false positive" nonsensical statements were given to four distinct LLMs (major columns). 6 questions were systematically queried across all prompt designs (major rows).Minor rows depict 20 different prompt designs, and minor columns the different input statements. Colour codings indicate comparison against a known "correct" answer, or if the query failed. Marginal distributions of answer correctness are shown on top (per statement) and left (per prompt design) in the heatmap cell .

🟦 Correct answer
□ Ambiguous answer/comparison
🟧 Wrong answer
🟥 Model prediction error or JSON parsing error
🟪 Input statements with nonsense or offtopic content

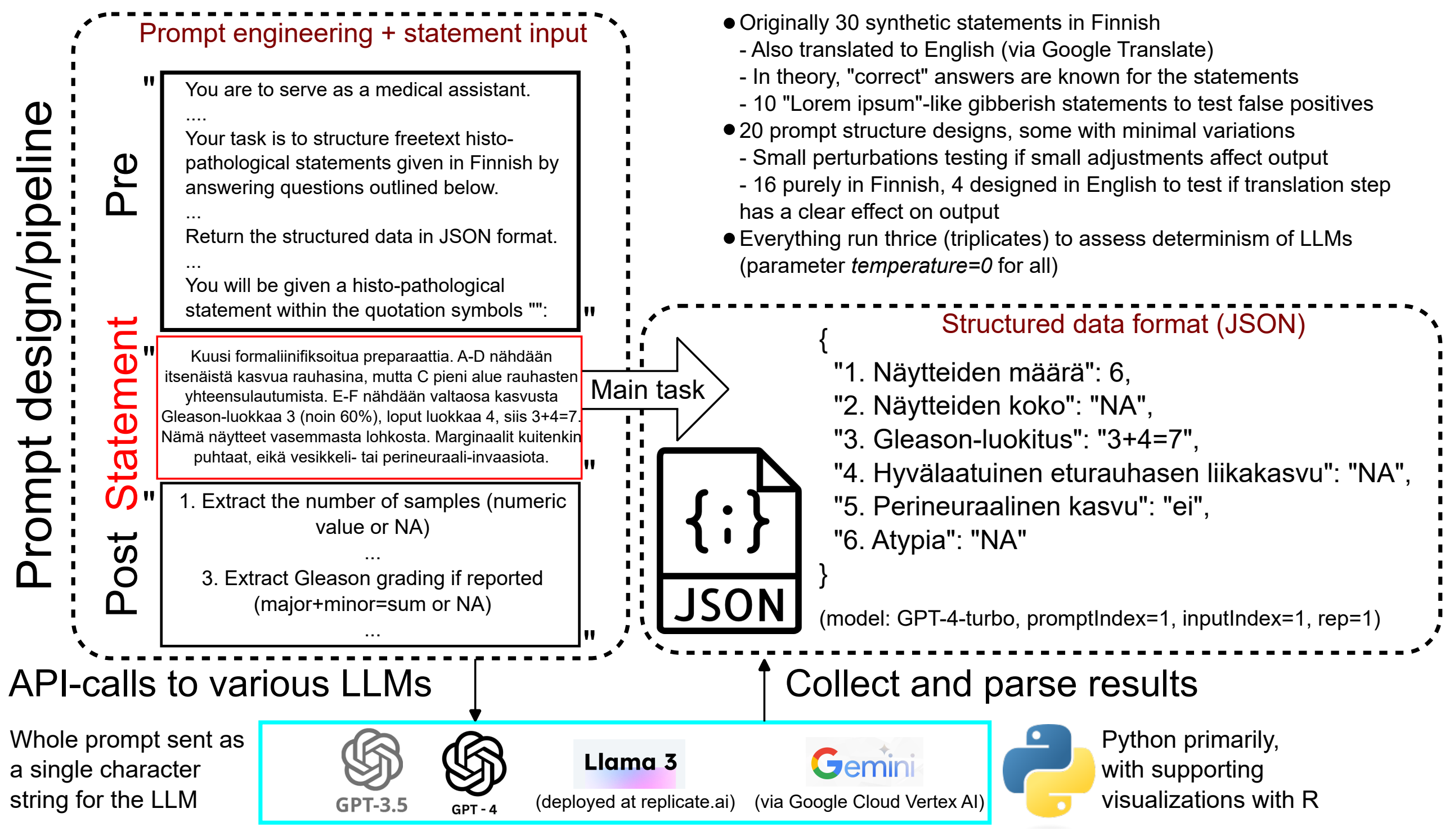
Introduction

Large Language Models (LLMs) have made significant breakthroughs across all scientific disciplines, including processing of biomedical text data. However, for specific domains such as patient confidentiality retaining electronic health records, usage of such models remains experimental and challenging. Deploying such models to real use cases is further hampered by limited availability of teaching data for smaller languages such as Nordic or Baltic languages.

In order to model data for research purposes and e.g. efficient visualizations, there is dire need for structuring freetext fields produced via multitude of electronic health care record pipelines. LLMs offer great potential in structuring freetext via simple structured data formats such as JSON, XML, or even tab-separated values. Here, we examine the use of three prominent LLMs for JSON structuring: GPT (both 3.5 and 4 releases), Gemini, and Llama 3. The utilized data comprises of 30 synthetic Finnish histo-pathological statements, as well as 10 non-sensical input statements testing LLM behaviour for "false positives". A total of 20 prompt designs were used, of which 16 were purely in Finnish and 4 in English coupled with the statements first machine-translated to English.

Materials & Methods

Figure 1 outlines the whole set of results collected via running a pipeline outlined in **Figure 2**. GitHub link for running the study: https://github.com/Syksy/HistPath_LLMpy


Prompt design/pipeline
Prompt engineering + statement input
Pre
Statement
Post
Main task
Structured data format (JSON)
JSON
API-calls to various LLMs
Collect and parse results
GPT-3.5
GPT - 4
Llama 3
Gemini

"You are to serve as a medical assistant.
....
Your task is to structure freetext histo-pathological statements given in Finnish by answering questions outlined below.
...
Return the structured data in JSON format.
...
You will be given a histo-pathological statement within the quotation symbols "":"

"Kuusi formaliinifiksoitua preparaattia. A-D nähdään itsenäistä kasvua rauhasina, mutta C pieni alue rauhasten yhteensulautumista. E-F nähdään valtaosa kasvusta Gleason-luokkaa 3 (noin 60%), loput luokkaa 4, siis 3+4=7. Nämä näytteet vasemmasta lohkosta. Marginaalit kuitenkin puhtaat, eikä vesikkeli- tai perineuraali-invaasiota."

"1. Extract the number of samples (numeric value or NA)
...
3. Extract Gleason grading if reported (major+minor=sum or NA)
..."

- Originally 30 synthetic statements in Finnish
  - Also translated to English (via Google Translate)
  - In theory, "correct" answers are known for the statements
  - 10 "Lorem ipsum"-like gibberish statements to test false positives
- 20 prompt structure designs, some with minimal variations
  - Small perturbations testing if small adjustments affect output
  - 16 purely in Finnish, 4 designed in English to test if translation step has a clear effect on output
- Everything run thrice (triplicates) to assess determinism of LLMs (parameter *temperature=0* for all)

```
{
 "1. Näytteiden määrä": 6,
 "2. Näytteiden koko": "NA",
 "3. Gleason-luokitus": "3+4=7",
 "4. Hyvälaatuinen eturauhasen liikakasvu": "NA",
 "5. Perineuraalinen kasvu": "ei",
 "6. Atypia": "NA"
}
```

(model: GPT-4-turbo, promptIndex=1, inputIndex=1, rep=1)

Whole prompt sent as a single character string for the LLM

(deployed at replicate.ai) (via Google Cloud Vertex AI)

Python primarily, with supporting visualizations with R

Figure 2: Rough outline for generating the results from prompt engineering to inspecting the newly structured data.

Results & Conclusions

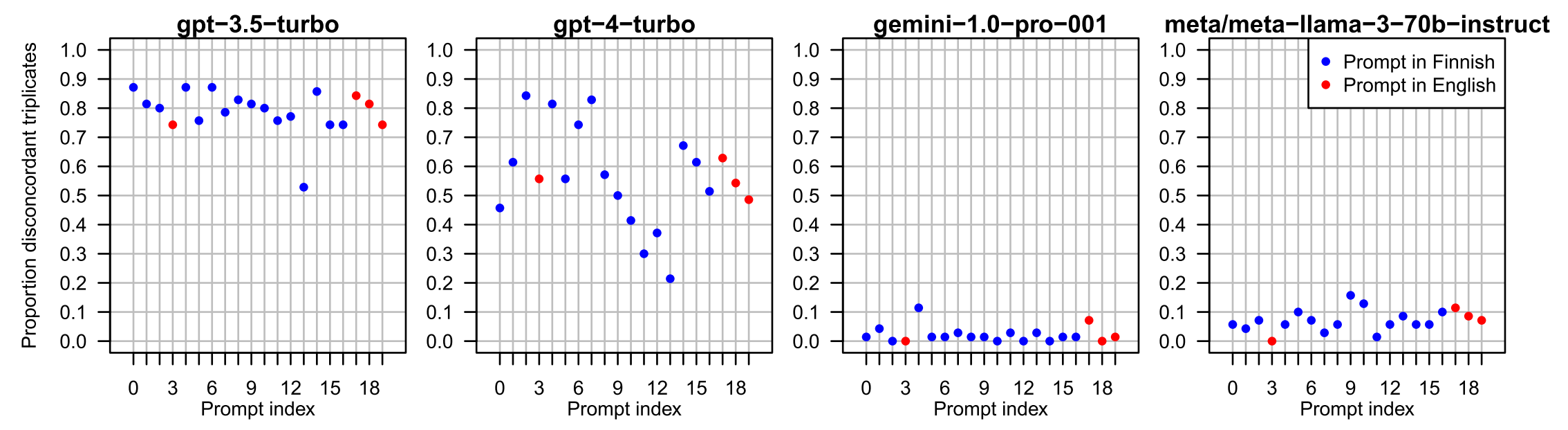


Figure 3: Substantial amount of disconcordance was noted across triplicates. A triplicate was considered disconcordant if even one of the three contained a single diffent character.

- Interestingly, as all runs were conducted as triplicates (with *temperature*-parameter set at 0 to promote determinism), a notable amount of variability was observed across the triplicates (**Figure 3**). This variability was notably highest in the GPT-family (both 3.5 and 4), but larger than zero also in Gemini and Llama. This highlights the fact that if exact scientific reproducibility is to be expected, using LLMs to structure data within a research pipeline poses a risk of producing stochastic differences across runs.
- Translation step to English (prompt indices 4, 17, 18, 19) had very minimal if any impact
- GPT-4 clearly outperformed GPT-3.5, particularly in avoiding "false positives" (pink minor columns in Figure 1), while both GPT-4 and Gemini presented stable performance. Llama 3 presented both excellent performance, but also severe inconsistencies for JSON structure (prompts 1, 9, 10, and 17)
- Even when set to minimal blockage, Gemini's harm/profanity filters prevented model predictions.

A summary of some key findings and differences is presented in **Table 1**.

Table 1: Brief summary of some of the key relative differences in the compared LLMs within the scope of the conducted empirical study.

	Pros	Cons
GPT-3.5	- Cheap and computationally effective - Native support for JSON structuring	- Clearly surpassed in performance by GPT-4 - Notable variability in returned responses across replicates - Closed source
GPT - 4	- Stable and competitive performance - Native support for JSON structuring	- Notable variability in returned responses across replicates - Closed source
Gemini	- Stable and competitive performance	- In EU/UK cannot be called directly with API-queries (req. Google Cloud Vertex AI environment) - Closed source - Issues with "harm/profanity" filters
Llama 3	- Closest to an open source alternative* - Potential for competitive performance	- Performance is highly sensitive to prompt design - Deviations from structuring instructions - License outlines "non-English" use as out of scope

*: Although often called open source, notable sources have disputed Llama's definition as open source

UNIVERSITY OF HELSINKI
UNIVERSITY OF TURKU
Prostate Cancer Research Group
HUS Helsinki University Hospital
JANE AND AATOS ERKKO FOUNDATION
Syöpäsäätiö Cancer Foundation Finland
F|C|I Finnish Cancer Institute
Cancer Society of Finland
SIGRID JUSELIUS STIFTELSE | SÄÄTIÖ | FOUNDATION
Finnish Cultural Foundation