Today's research for tomorrow's health

syreon

Research Institute

Comparative Analysis of Large Language Models for Extracting Patient-Reported Outcome Measures from Clinical Trial Protocols in Lymphoma

Attila Imre, Dalma Hosszú, Anna Bogos, Balázs Nagy, Judit Józwiak-Hagymásy, Tamás Ágh, Ákos Bernard Józwiak

Challenges of PROM Extraction from Clinical Trial Protocols

- Patient Reported Outcome Measures (PROMs) are critical for regulatory & HTA decisions
- Challenges from a data extraction standpoint:
  - Inconsistent terminology
  - Manual curation is slow
- Lymphoma trials apart from large patient relevance, utilize an array of generic and specific PROMs making it ideal for experimentation.

Aim

Evaluate zero-shot LLMs (gemma2, llama3.3-70b, gpt-4o-mini) against an expert gold standard dataset in a sample of lymphoma trial protocols

Dataset Curation

- A search through the official API of ClinicalTrials.gov for the condition "*lymphoma*"* yielded 9,378 trials on 2024 November

- Out of the total number of trials, we selected Phase III (722) and Phase II+III (85) trials, this totalled 807 trials

- We combined primary, secondary and other outcomes, calculated their **lengths and used a weighted sampling procedure to select 300 trials**.

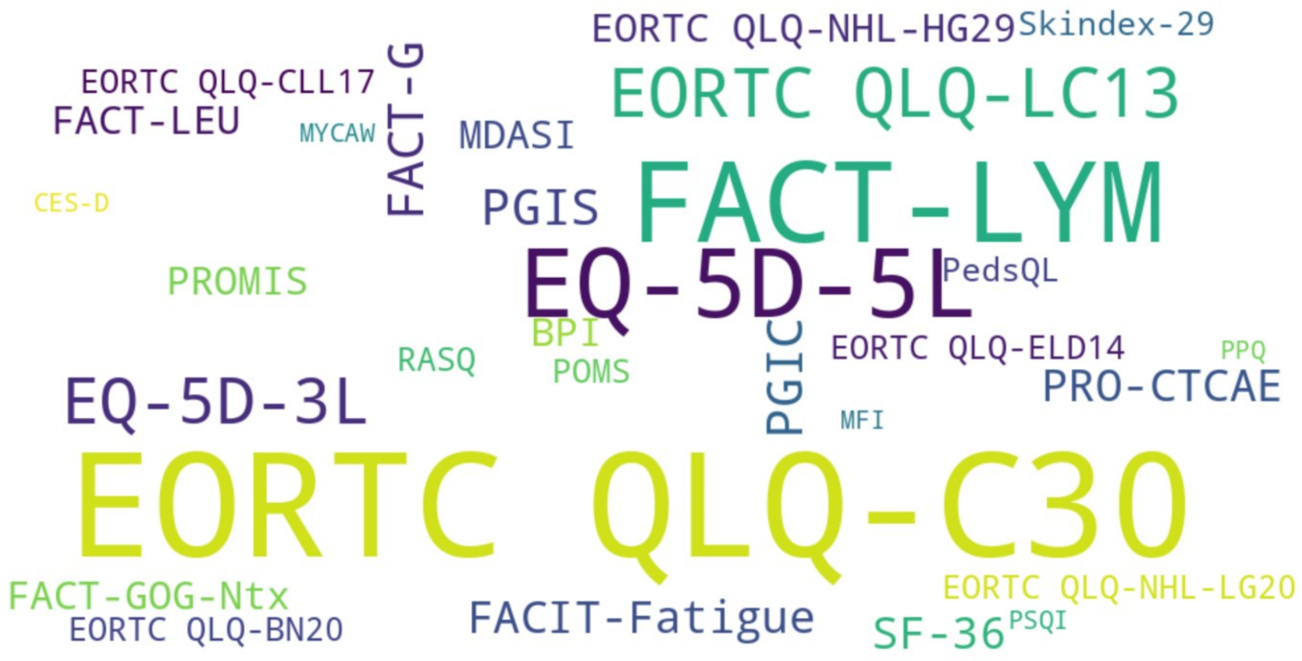
* With synonyms: Lymphoma; Lymphomas; Malignant lymphoma; Malignant Lymphomas; Lymphomatous; Lymphosarcoma

Dataset Curation

- Four reviewers with domain expertise (DH, AB, ÁT, ÁJ) extracted all the PROMs, which was double check by another reviewer (AI)
- **107 trials (35.67%)** contained at least one PROM, 193 trials (64.33%) did not contain PROMs, making our dataset imbalanced
- The mean (±SD) number of PROMs was 0.80 ± 1.38
- One trial contained up to 10 PROMs

PROMs within the dataset

PROM	Count
EORTC QLQ-C30	60
FACT-Lym	33
EQ-5D-5L	24
EORTC QLQ-LC13	10
EQ-5D-3L	9
PGIS	5
FACT-G	5
PGIC	5
SF-36	4
FACIT-Fatigue	4



LLMs and Extraction Pipeline

We tested the capabilities of distinct LLMs, namely:

1. Gemma-2-9b-it: 9 B parameter, open-source (Google)
2. Llama3.3-70b-it-turbo: 70 B parameter, open-source (Meta)
3. GPT-4o-mini-2024-07-18: Proprietary “Omni” model (OpenAI)

- Custom zero-shot prompt for PROM identification
- Queries were issued via LLamaIndex v0.11.20
    - Gemma & Llama models through DeepInfra.com
    - GPT-4o-mini through OpenAI’s official API
- All experiments were conducted using Python v.3.9.13.

Sample Input & Output

Sample Input

Outcome: Global health status/quality of life (GHS/QoL), Up to 5 years from the last participant randomized
Description: Time from randomization to first confirmed clinically meaningful improvement from baseline in the European Quality of Life Module Chronic Lymphocytic Leukemia 17 (EORTC QLQ-CLL17)

Sample Output

PROOutput(pros=[PRO(abbreviation='EORTC QLQ-CLL17')])

Evaluation Framework

- Unit of analysis: each (Trial, PROM) pair
- Metrics:
  - True Positive (TP) / False Positive (FP) / False Negative (FN) / True Negative (TN) at PROM level
  - Precision, Recall, F1, Accuracy
- Statistical test: McNemar's test

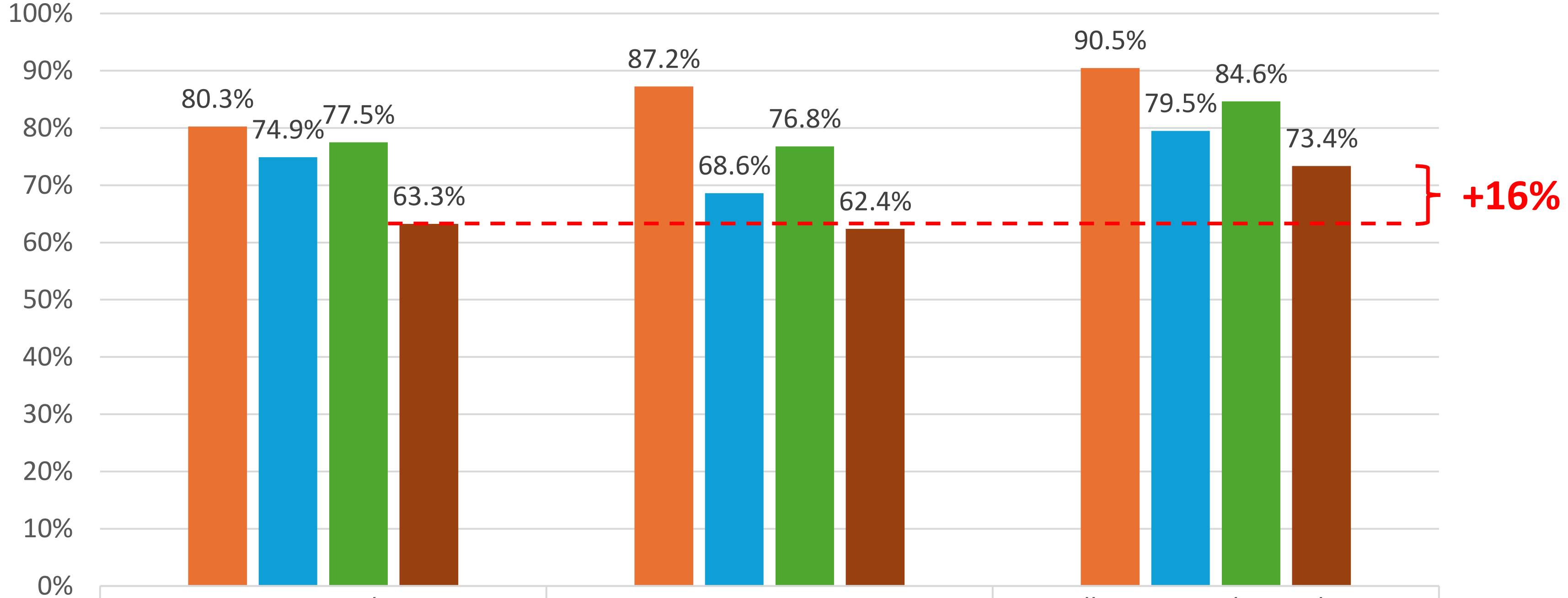
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Comparative Analysis of LLMs I.



	gemma-2-9b-it	gpt-4o-mini-2024-07-18	llama-3.3-70b-it-turbo
Precision (%)	80.3%	87.2%	90.5%
Recall (%)	74.9%	68.6%	79.5%
F1-score (%)	77.5%	76.8%	84.6%
Accuracy (%)	63.3%	62.4%	73.4%

Key takeaways

- Baseline accuracy is 63.3 % (gemma-2-9b-it).
- Llama3.3-70b-it-turbo raises accuracy to 73.4 %, a ~16 % relative improvement over baseline.
- GPT-4o-mini (62.4 %) performs on par with gemma (-1.4 % relative).

syreon Research Institute

ISPOR 2025

Comparative Analysis of LLMs II.

Model A	Model B	Model A is correct, and Model B is wrong	Model A is wrong, and Model B is correct	OR	95 % CI	χ^2	p-value
gemma2-9b-it	llama3.3-70b-it-turbo	20	55	8.24	[4.52, 15.03]	15.41	< 0.001
gemma2-9b-it	gpt-4o-mini-2024-07-18	54	63	1.82	[1.10, 3.00]	0.55	0.46
llama3.3-70b-it-turbo	gpt-4o-mini-2024-07-18	52	26	5.98	[3.36, 10.63]	8.01	< 0.05

Llama3.3-70b-it-turbo compared to…

- Gemma-2-9b-it; OR = 8.24 (95 % CI 4.52-15.03, $p < 0.001$)
    - When the two models disagree, Llama is over eight times more likely than Gemma to correctly extract a PROM.
- GPT-4o-mini; OR = 5.98 (95 % CI 3.36-10.63, $p < 0.05$)

Trial Coverage

- Gold-standard trials with ≥1 PROM: 107 trials
- Trials with ≥1 predicted PROM:
  - gemma2-9b-it: **111**
  - gpt-4o-mini-2024-07-18: **104**
  - llama3.3-70b-it-turbo: **102**

Llama3.3-70b-it-turbo was the most conservative (fewer false positives of empty trials)

- When EORTC QLQ-C30 was within the PROMs llama3.3 found it 100% of the times, gemma 93% of the time, gpt-4o-mini 97% of the time.

Which PROMs were most difficult to find?

PROM	Total Errors	False Positive Count	False Negative Count
EQ-5D-3L	30	0	30
EQ-5D	28	28	0
EQ-5D-5L	15	11	4
FACT-G	14	8	6
PGIS	13	0	13
PGIC	12	0	12

Patterns:

- Version confusion (EQ-5D variants)
- Inclusion of FACT-G when nested inside FACT-LYM
- PGIC/PGIS were missed

Abbreviations: PGIS: Patient Global Impressions of Severity; PGIC: Patient Global Impressions of Change

Conclusions

- The best performing LLM was llama3.3-70b-it-turbo (Accuracy 73.4 %, Precision 90.5 %, Recall 79.5 %) on this limited dataset.
- Use cases:
  - Pre-screening for systematic reviews
  - Live evidence tracking of PROM usage trends
- Future work:
  - Automated normalization of PROM variants
  - Build (semi-) automatic verification frameworks around LLM predictions based on simpler (and more interpretable) models.

syreon
Research Institute

H-1142 Budapest Mexikói str. 65/A
Website: http://www.syreon.eu

Telephone : +36 1 787 0083
Email : info@syreon.eu