PhenoAgent – Agent-based architecture for high-throughput deep phenotyping with Large Language Models

Marek Wiewiórka[1], Wojciech Sitek[1], Rafał Małanij, Tomasz Gambin[1]
[1] Institute of Computer Science, Warsaw University of Technology, marek.wiewiorka@pw.edu.pl

Background

Deep phenotyping refers to the comprehensive and detailed analysis of phenotypic traits in organisms to understand complex biological processes and diseases. **Human Phenotype Ontology** (HPO) that is one of the most popular ontologies for computational phenotype analysis currently contains over 18,000 terms and over 156,000 annotations to hereditary diseases. Over the years a number of automatic methods has been developed, such as rule-based and machine learning, including recent evaluations of **Large Language Models** (LLMs) applicability ([7, 1]). **PhenoAgent** is, to the best of our knowledge, the first LLM-based tool for an automatic HPO terms tagging that relies on **Retrieval Augmented Generation** (RAG)[3] and **Mixture-of-Agents** (MoA)[6] concepts.

Methods

Our solution implements deep phenotyping process using **DSPy** framework combined with **Sentence Transformers** library for HPO embeddings model fine-tuning and **LanceDB** for hybrid search of HPO terms in RAG component. LLMs are exposed using **OpenAI** protocol (for quantized models with 8-14B parameters using Ollama, Llama3.1 70B and 405B in the Azure Cloud). **NeuML/pubmedbert-base-embeddings** was fine-tuned using **MultipleNegativesRankingLoss** loss function. PhenoAgent's user interface and application programaming interface are implemented with **Gradio** library. PhenoAgent can be deployed locally, on-premise and in the cloud environments.

References

[1] T. Groza, H. Caufield, D. Gration, G. Baynam, M. A. Haendel, P. N. Robinson, C. J. Mungall, and J. T. Reese. An evaluation of GPT models for phenotype concept recognition. *BMC Medical Informatics and Decision Making*, 24(1):30, Jan. 2024.

[2] T. Groza, D. Gration, G. Baynam, and P. N. Robinson. FastHPOCR: pragmatic, fast, and accurate concept recognition using the human phenotype ontology. *Bioinformatics*, 40(7):btae406, July 2024.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

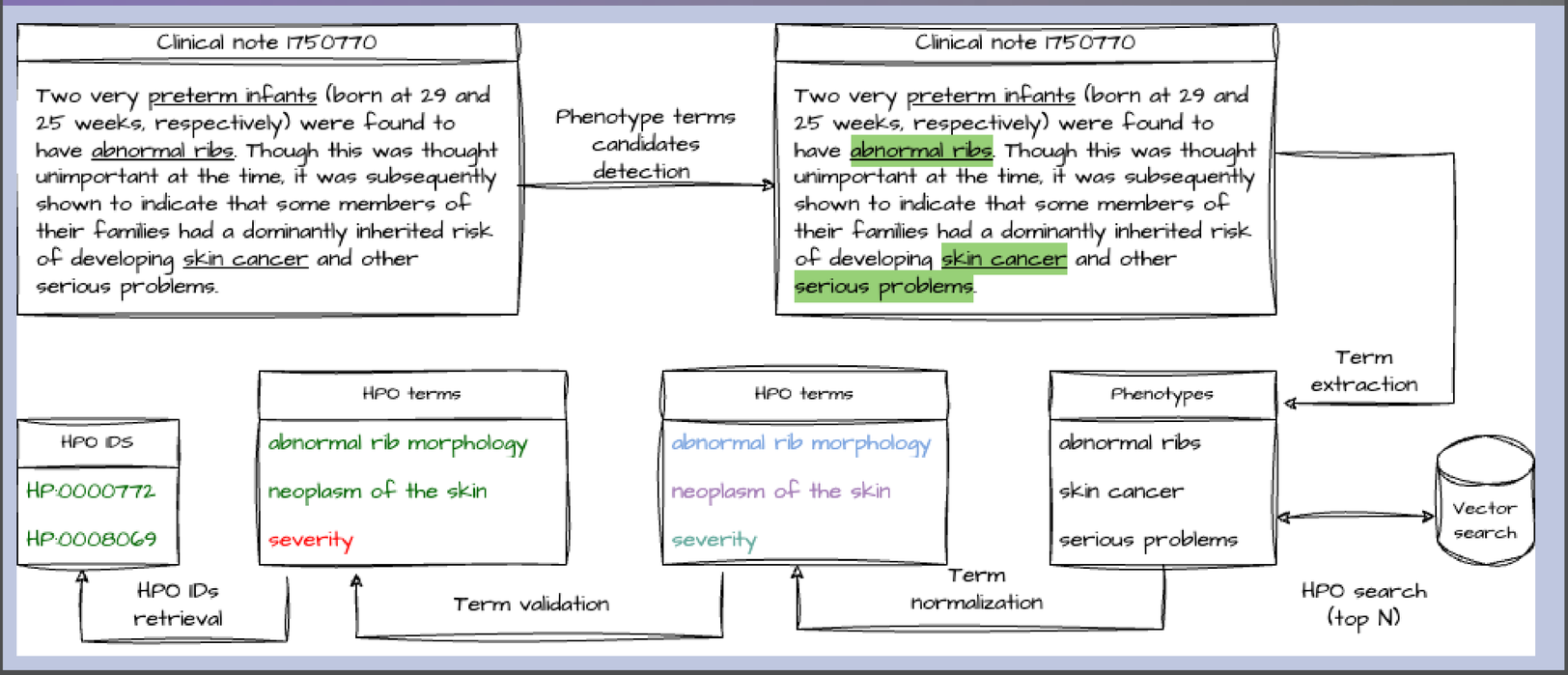
[4] M. Lobo, A. Lamurias, and F. M. Couto. Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules. *BioMed Research International*, 2017:1–8, 2017.

[5] L. Luo, S. Yan, P.-T. Lai, D. Veltri, A. Oler, S. Xirasagar, R. Ghosh, M. Similuk, P. N. Robinson, and Z. Lu. PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, 37(13):1884–1890, July 2021.

[6] J. Wang, J. Wang, B. Athiwaratkun, C. Zhang, and J. Zou. Mixture-of-Agents Enhances Large Language Model Capabilities, June 2024.

[7] J. Yang, C. Liu, W. Deng, D. Wu, C. Weng, Y. Zhou, and K. Wang. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. *Patterns*, 5(1):100887, Jan. 2024.

Deep phenotyping process



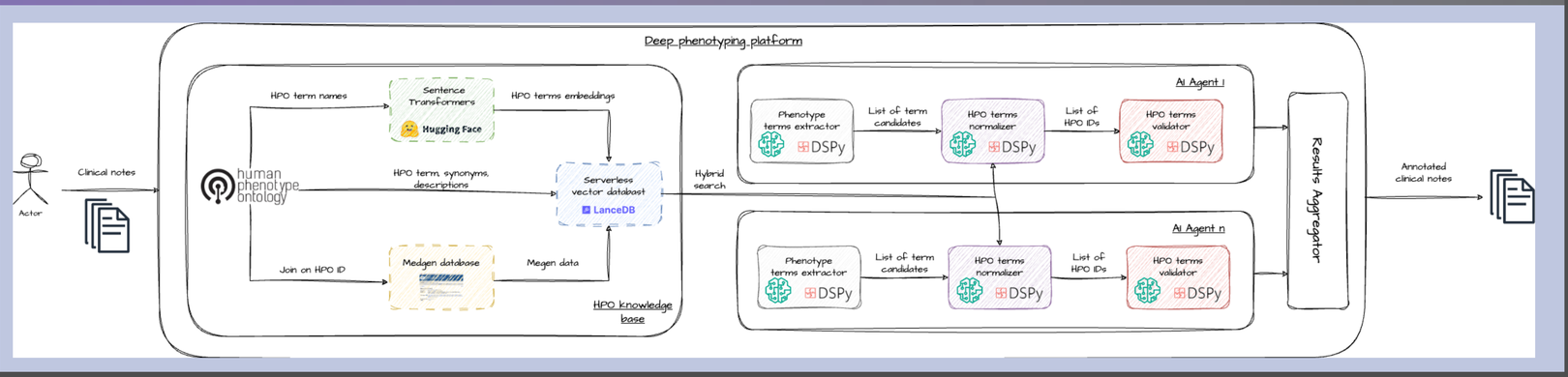
Mixture-of-Agents (MoA) voting strategy

Let k be the minimum number of models required for an element to be included in the final output, M be the number of models, S_i be the set of predicted elements from the i-th model where $S_i \subseteq \mathcal{U}$ and $\mathcal{U}$ is the universal set of all possible elements.

$$f(x) = \sum_{i=1}^{M} \mathbf{1}(x \in S_i)$$
$$S_{\text{MoA-M-k}} = \{x \in \mathcal{U} \mid f(x) \geq k\} \quad (1)$$

where $\mathbf{1}(x \in S_i)$ is an indicator function that is 1 if x is in S_i, and 0 otherwise.
PhenoAgent-MoA-M-k is the identifier of the strategy presented in the results section.

PhenoAgent architecture



Results

Tool	Model	Precision	Recall	F1
PhenoGPT[7]	Llama2-7B	0.3136	0.2805	0.2961
PhenoTagger[5]	BioBert	**0.7992**	0.6971	**0.7447**
FastHPOCR[2]	-	0.6503	**0.7303**	0.6880
PhenoAgent-MoA-8-2	MoA-8[a]	0.4974	0.6990	0.5812
PhenoAgent-MoA-8-3	MoA-8	0.6275	0.6241	0.6258
PhenoAgent-Llama-70	LLama3.1-70B	0.5549	0.5549	0.5401
PhenoAgent-Llama-405	LLama3.1-405B	0.6248	0.5616	0.5915

Performance comparison of different tools using a 10% subset of BiolarkGSC+[4] dataset. Bold to indicate the highest values, colors for LLM-based solutions comparison.

[a] MoA architecture using 8 LLMs:, 8 and/or 4-bit quantization: Llama3.1-8B(4,8), Gemma2:9B(4,8), Phi3:14b(4), Hermes3:8B(4,8). Mistrsal:7B(4).

Discussion

1. **PhenoAgent** is a prototype project that already achieves comparable results with other **state-of-the-art** tools on the GSC+ corpus (i.e. PhenoTagger or FastHPOCR) without **any LLM fine-tuning**.
2. It proves superiority of **RAG** architectures for HPO concept normalization (i.e.when compared to PhenoGPT) and help to avoid HPO ID **hallucinations**([7]).
3. It shows that the **MoA** architecture of relative small LLMs can improve inference performance and outperform state-of-the-art models, e.g. **LLama-3.1-405B**.
4. Further work on a hybrid (dictionary + LL) solution for addressing lower than expected inference performance, e.g. using DSPy **optimizers**.
5. **Polish** language support is on the project roadmap.