End-to-End Deep Learning for Named Entity Recognition and Relation Extraction in Gut-Brain Axis PubMed Abstracts

Aleksis Datseris **Mario Kuzmanov** **Ivelina Nikolova-Koleva** **Dimitar Taskov** **Svetla Boytcheva**

Gut-BrainIE Task at CLEF 2025 BioASQ Lab

The challenge aims to promote automatic extraction of gut-brain related facts from scientific literature and evaluates 4 subtasks across 13 named entity categories and 19 relation predicates:

- Named Entity Recognition (NER)
- Binary relation extraction (Binary RE)
- Ternary tag-based relation extraction (Tag RE)
- Mention-based relation extraction (Mention RE)

Named Entity Recognition

- GatorTron-Base - pretrained on clinical notes, fine-tuned on $Platinum \cup Gold \cup Silver \cup Bronze$.
- GLiNER - pretrained on the PileNER corpus, fine-tuned on $Platinum \cup Gold \cup Silver \cup Bronze \cup BronzeAugmented$.
- BNLP ELECTRA - re-trained on PubMed abstracts.
- ScispaCy - `en_core_sci_md`, fine-tuned only on $Platinum \cup Gold \cup Silver$.
- Gazetteer matching with exact match.
- LLM approach - GPT-4.1 (3-shot) with examples from $Platinum$ and (0-shot) with description for each of the 13 categories.

Named Entity Recognition Results

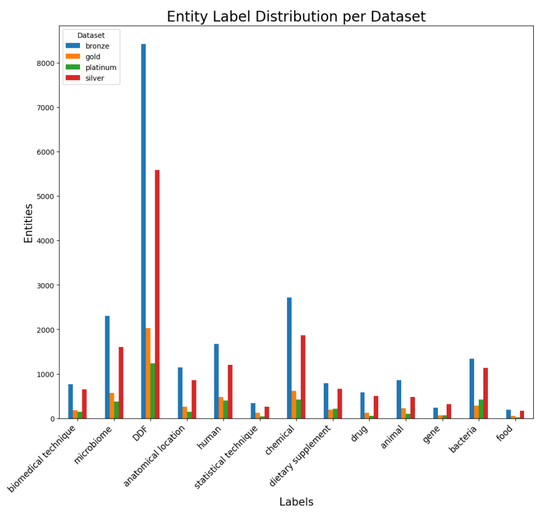
Model	P	R	F1
GatorTron-Base	77.37%	82.63%	79.91%
BNLP ELECTRA	82.60%	82.45%	**82.53%**
GLiNER	81.05%	**82.72%**	81.88%
ScispaCy	76.65%	71.71%	74.10%
Gazz	16.30%	12.71%	14.29%
ScispaCy + Gazz + BNLP ELECTRA	**91.41%**	69.56%	79.00%
ScispaCy + Gazz	56.38%	70.37%	62.60%
GPT (3-shot)	45.88%	67.32%	54.57%
GPT (0-shot)	42.01%	54.79%	46.38%

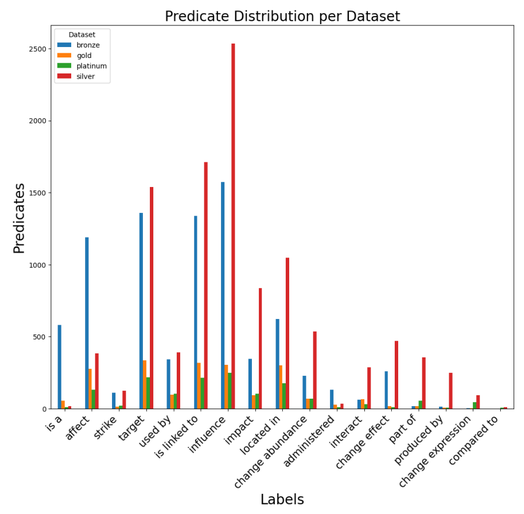
Acknowledgements

RES-Q+ Activate Quality

HEREDITARY

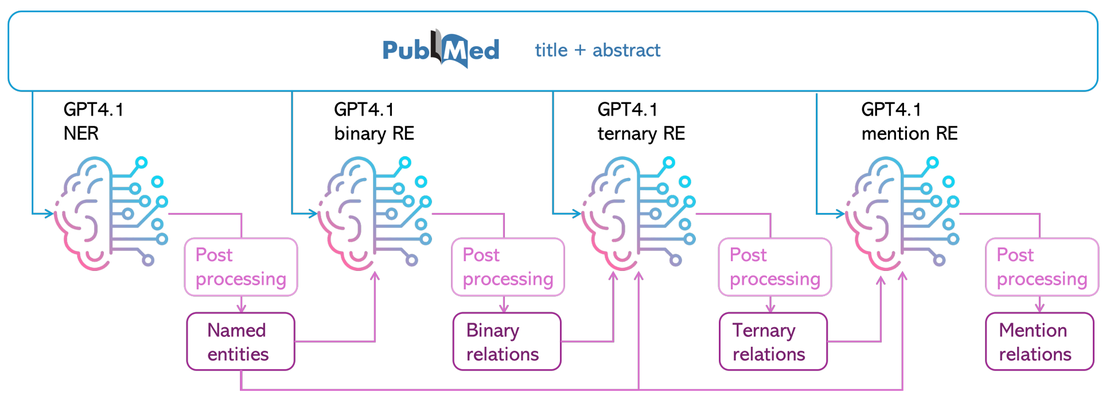
Gut-Brain Datasets





Relation Extraction Approaches

- ATLOP - supervised powerful and efficient relation extraction method that handles multi-label classification of relations through automatic thresholding.
- LLM approach - GPT 4.1 (0-shot)(1-shot)(2-shot):



Relation Extraction Results

Model	P	R	F1	Task
KRISSBERT	68.04%	60.00%	63.77%	Binary RE
GPT (1-shot)	57.99%	**70.91%**	**63.80%**	Binary RE
KRISSBERT	67.35%	57.39%	61.97%	Tag RE
GPT (2-shot)	**95.10%**	**84.35%**	**89.40%**	Tag RE
KRISSBERT	46.60%	35.54%	40.32%	Mention RE
GPT (1-shot)	29.01%	**44.29%**	35.05%	Mention RE

Discussion, Error Analysis, Conclusions

- Most challenging categories - "gene", "chemical".
- Models performed well on the span detection but struggled on the category classification task.
- ATLOP performs well on binary and tag-based RE and struggles on mention-based RE.
- Predicates "impact", "influence", and "affect" are often mistaken with "is linked to".
- LLMs demonstrate very good performance on binary and tag-based RE and struggle to identify correct mentions, and a common issue is the mismatch of the relation direction.