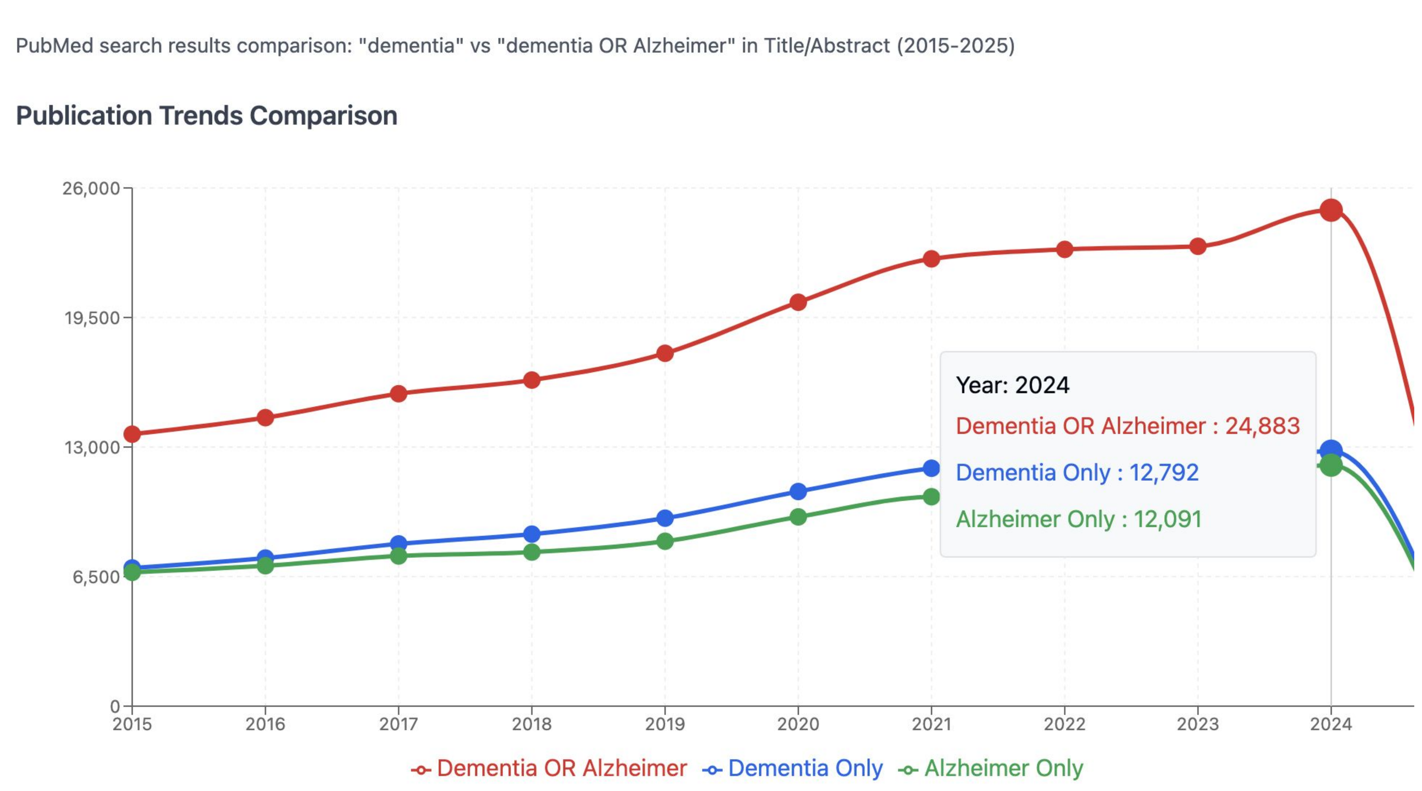
A Large Language Model based Framework for Dementia Related Hypothesis Generation

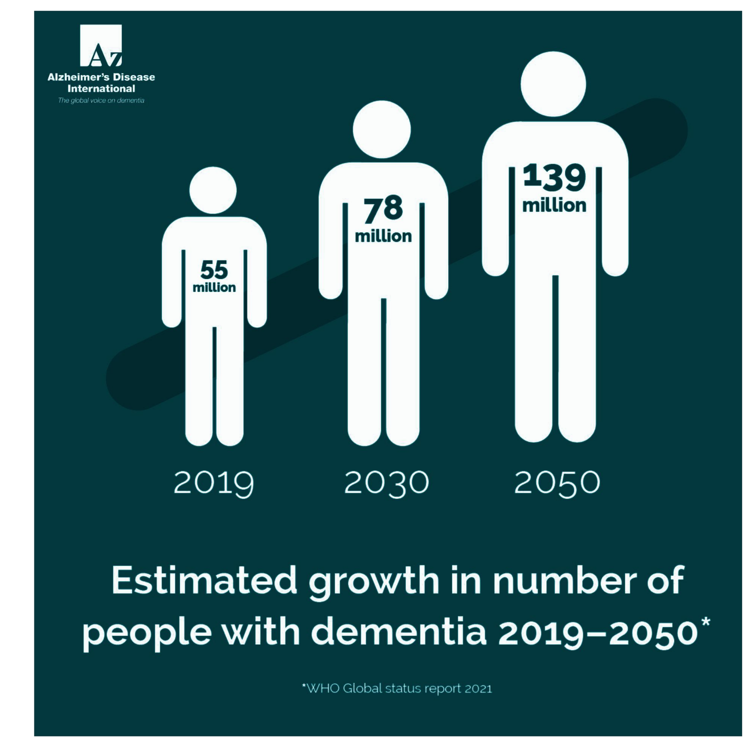
Rosni Vasu[1], Hang Dong[2,*], Yusuf Abdulle[3], Judith Harrison[4], and Honghan Wu[5]

[1]University of Zurich, Switzerland; [2]University of Exeter, Exeter; [3]King's College London, London; [4]Newcastle University, Newcastle upon Tyne, England, United Kingdom; [5]University of Glasgow, Glasgow, Scotland, United Kingdom

Motivation



- Dementia is a progressive neurological disorder affecting millions worldwide, posing significant challenges for healthcare and research.
- Huge amount of scientific literature makes it hard to keep track of relevant information and new hypotheses.
- Large Language Models (LLMs) show great promise for hypotheses generation in certain fields, but remain under-explored for dementia domain.
- Integrating structured biomedical knowledge graphs (KGs) could "inspire" LLMs and boost novelty, relevance and testability of generated hypotheses.



Structured Dataset Curation

Despite the explosion in dementia literature, hypothesis generation remains slow, manual, and fragmented. Our approach leverages **LLMs + domain-specific knowledge graphs** to accelerate and enhance the discovery of novel and feasible insights -- critical for guiding future studies such as identifying modifiable risk factors.

"components": {
    "independent_variable": "High GFAP levels",
    "dependent_variable": "risk of developing other types of dementia"
  },
  "relation": {
    "relation": "increases"
  }
}

Sentence: "High GFAP levels were also associated with an increased risk of developing other types of dementia (HR = 3.02, 2.21, and 3.05 for ADRD, VD, and FTD, respectively)"
Context: "..."

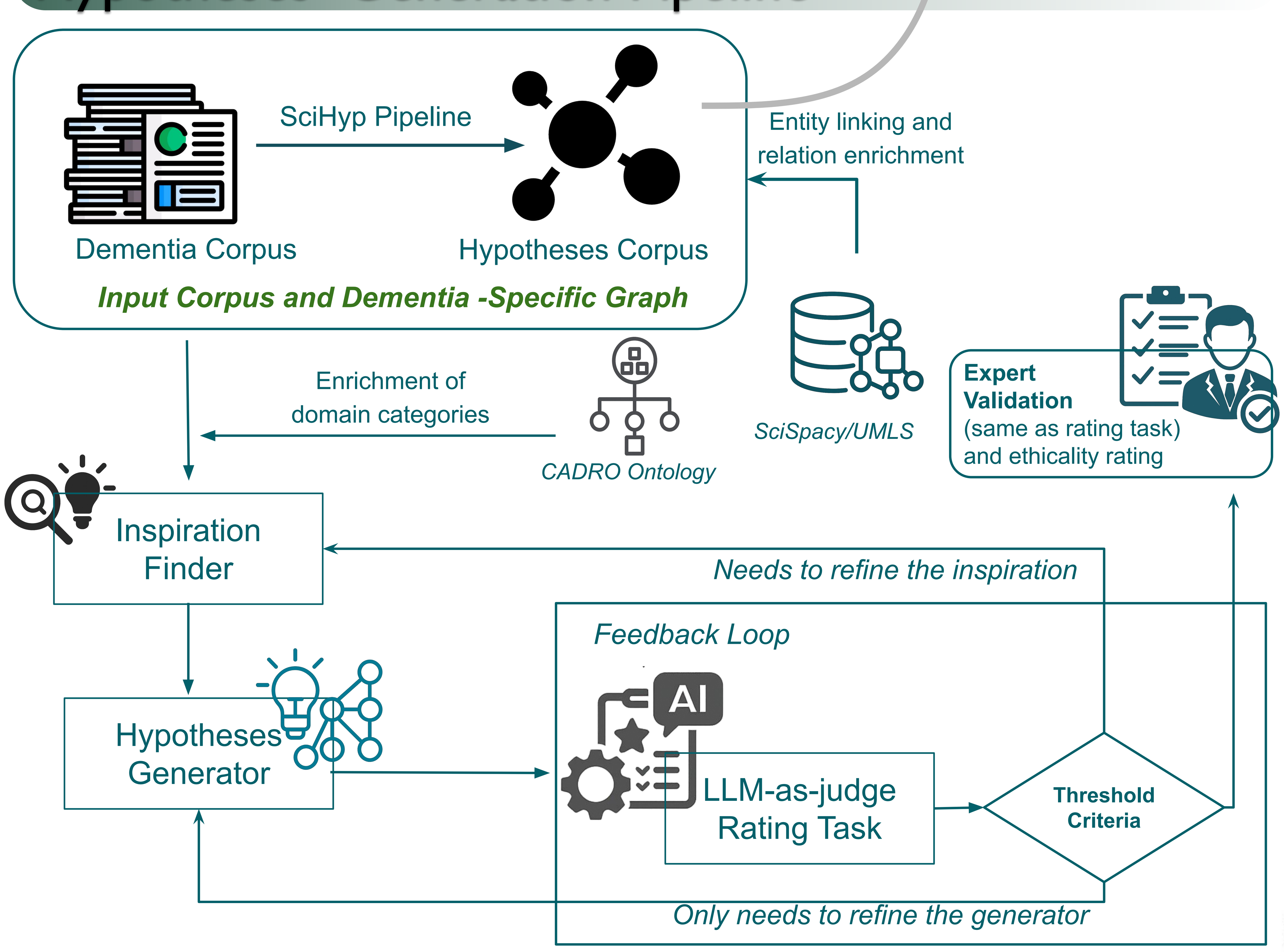
Sentence: "These researchers found that, when controlling for age, on average, \"high PA\" mutation carriers developed very mild dementia an average of 15 years later than those in the \"low PA\" group"
Context: "..."

"components": {
    "left_group": "high PA mutation carriers",
    "right_group": "low PA mutation carriers",
    "variable": "age of onset of very mild dementia"
  },
  "relation": {
    "operator": "hasLaterValueThan"
  }
}

Examples from Hypotheses Corpus

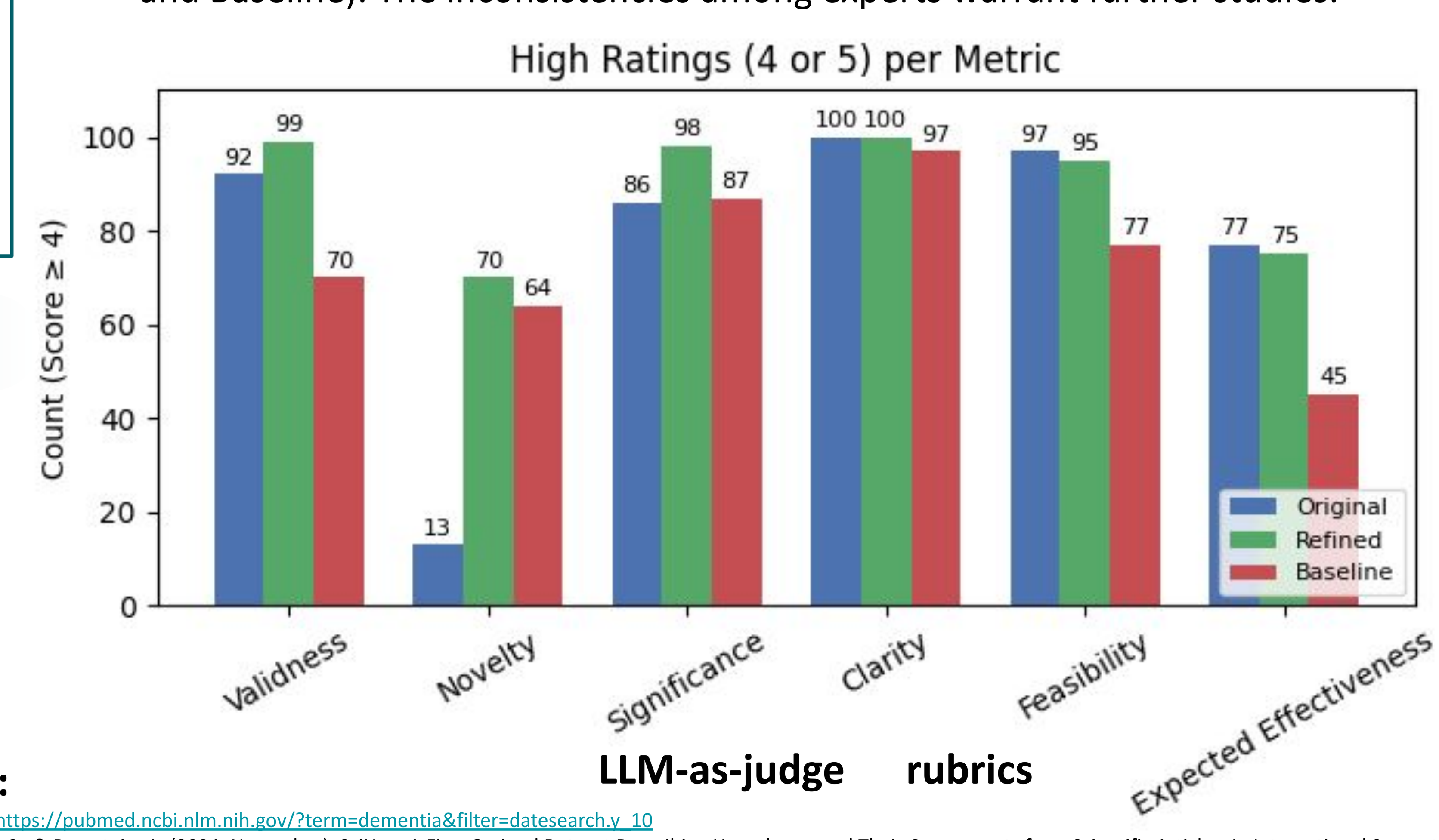
Number of papers	60
variables with left_group - right_group relationships	978
Unique independent_variables, dependent_variables	3281, 3430
Unique left_group, right_group	3116, 3260
Unique relations	89

Hypotheses Generation Pipeline



Domain-specific structured knowledge integration

- Extracted hypotheses components were linked to UMLS concepts to enable graph-level representation and exploration.
- We enriched extracted hypotheses by aligning variables to CADRO categories, a classification system to help identify research gaps in dementia, supporting structured navigation and hypothesis expansion.
- **Evaluation strategy: (1)** Baseline: 10 abstracts × 10 hypotheses = 100 (Solely GPT-4o) **(2)** Ours (backbone GPT-4o): 10 variable sets × 10 hypotheses. 200 in total (100 Original and 100 Refined). **LLM-as-judge results see Figure below.**
- **Expert Review:** Blind evaluation of 50 hypotheses, 25 each for the baseline and Ours (refined) hypotheses using 6 criteria (e.g., Validness, Novelty, Effectiveness), adapted from Yang et al. (2024) with 2 experts (a senior expert as Clinical lecturer, and a junior expert as PhD student).
- **Expert Evaluation Results:** The senior expert scores **higher in all 6 criterias**, a macro-average of 86% of the hypotheses with scores >= 4 for the Refined vs. 38% for the Baseline. The junior expert scores similarly (54% for both Refined and Baseline). The inconsistencies among experts warrant further studies.



LLM-as-judge rubrics

Discussion

- **Expert evaluation** shows the advantage of our approach using refinement using LLM feedback and Knowledge Graph. **LLM-as-judge evaluation** suggests our approach notably improves novelty of hypotheses, but slightly reduces feasibility and expected effectiveness.
- **Ongoing/Future studies:** (1) Qualitative analysis of generated hypotheses with detailed expert analysis and rating; (2) Extended expert evaluation with larger samples;

  (3) Experiments with more ablation settings, e.g., on the role of KGs (UMLS and CADRO);

  (4) Advancing LLMs to extract structured hypotheses from papers, instead of using BERT.

* Main presenter: H.Dong2@exeter.ac.uk

References:

1. Pubmed search: https://pubmed.ncbi.nlm.nih.gov/?term=dementia&filter=datesearch.y_10
2. Vasu, R., Sarasua, C., & Bernstein, A. (2024, November). SciHyp: A Fine-Grained Dataset Describing Hypotheses and Their Components from Scientific Articles. In International Semantic Web Conference (pp. 134-152). Cham: Springer Nature Switzerland.
3. Yang, Zonglin, et al. "Large Language Models for Automated Open-domain Scientific Hypotheses Discovery." Findings of the Association for Computational Linguistics ACL 2024. 2024.
4. Alzheimer's Disease International. World Alzheimer Report 2024. https://www.alzint.org/resource/world-alzheimer-report-2024/
5. Refolo, L. M., Snyder, H., Liggins, C., Ryan, L., Silverberg, N., Petanceska, S., & Carrillo, M. C. (2012). Common Alzheimer's Disease Research Ontology: National Institute on Aging and Alzheimer's Association Collaborative Project. Alzheimer's & Dementia, 8(4), 372–375. (CADRO Ontology: https://iadrp.nia.nih.gov/about/cadro)