Evaluating LLMs' knowledge using word senses and general knowledge quizzes

Anna Mosolova

LAMSADE, PSL & ALMAnaCH, INRIA

About me

- MSc in NLP at Université de Lorraine
  - Thesis topic: Machine translation of low-resource languages of Eurasia
- PhD in NLP at Université Paris Cité
  - Thesis topic: Semi-supervised word sense and frame induction
- Postdoc at ALMAnaCH team in INRIA Paris
  - Topic: LLM evaluation in specialized domains (finance, education, healthcare, public administration, culture)
- AI Fellow at LAMSADE and ALMAnaCH
  - Topic: LLM evaluation in specialized domains (finance, education, healthcare, public administration, culture)

Presentation outline

1. Can LLMs distinguish between different senses of the same word?
   - The task: Word Sense Induction
   - Different approaches: clustering vs. direct prompting of LLMs
   - Improvements to the clustering-based approach
2. How well do LLMs perform on quizzes?
   - Quiz time
   - LLM performance across (i) languages, (ii) topics, (iii) difficulty levels, and (iv) time periods
   - Improvements using agent mode
3. Conclusions and future work

Word Sense Identification Evaluation

Word senses

Example: *Anna a* **détaché** *l'étiquette de sa chemise*

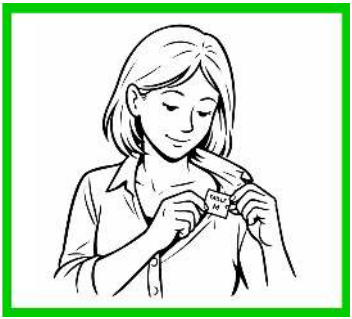
What is the sense of **détacher**?

Translation: Anna detached the label from her shirt

Word senses

Example: *Anna a* **détaché** *l'étiquette de sa chemise*

What is the sense of **détacher**?



Translation: Anna detached the label from her shirt

How can we automatically distinguish different word senses?

Two main tasks:

- Word Sense Disambiguation
  - Supervised
- Word Sense Induction
  - Unsupervised

How can we automatically distinguish different word senses?

Two main tasks:

- Word Sense Disambiguation
  - Supervised
- Word Sense Induction
  - Unsupervised

Word Sense Disambiguation:

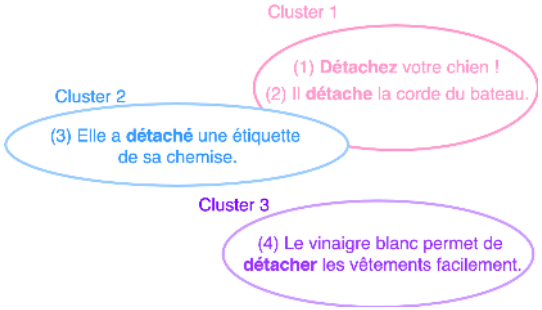
Example	Sense
(1) **Détachez** *votre chien !*	untie
(2) Elle a **détaché** *une étiquette de sa chemise.*	tear off
(3) Le vinaigre blanc permet de **détacher** *les vêtements facilement.*	unstain
(4) Il **détache** *la corde du bateau.*	???

How can we automatically distinguish different word senses?

Two main tasks:

- Word Sense Disambiguation
  - Supervised
- Word Sense Induction
  - Unsupervised

Word Sense Induction:



How can we automatically distinguish different word senses?

Two main tasks:

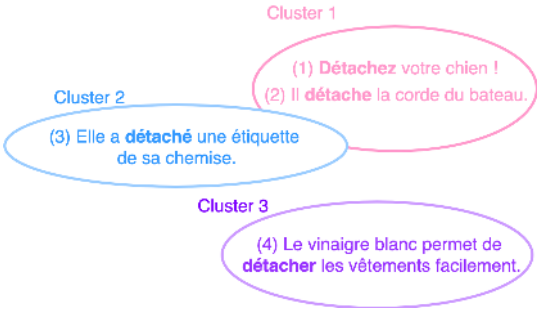
- Word Sense Disambiguation
  - Supervised
- Word Sense Induction
  - Unsupervised

→ Word Sense Induction doesn't require any training data, so we will use it!

Approaches towards Word Sense Induction

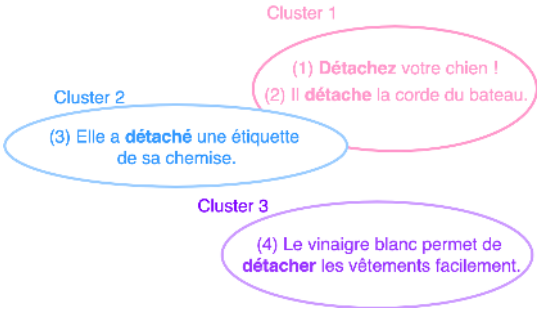
Approaches towards Word Sense Induction

Clustering

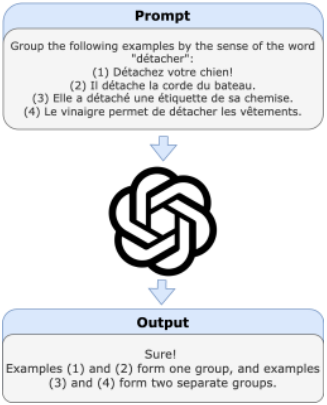


Approaches towards Word Sense Induction

Clustering



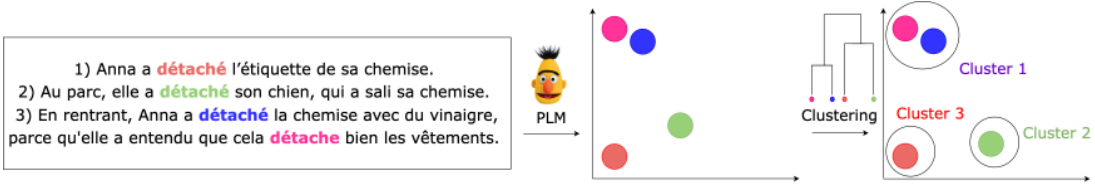
Large Language Models



Clustering-based Word Sense Induction

Two main components:

- Vectorizer for the target words
- Clustering algorithm



Evaluated models

- **Baseline (clustering):**
  - 1cpl: One cluster per lemma (word)
- **Previous best models:**
  - PolyLM (Ansell et al., 2021)
    - Sense embeddings from BERT (attributed to one sense by construction)
  - LSDP (Amrami et al., 2019)
    - Clustering vectors of most probable substitutions of a target word from BERT
- **Direct prompting of LLMs:**
  - GPT-4o
  - Llama 3.1 8B Instruct
  - Llama 3.3 70B Instruct
- **Our model (clustering):**
  - BERT-large target word embeddings + Agglomerative clustering (w/ silhouette score)

Evaluation dataset and metric

Dataset: SemCor-WSI

- 15,308 examples for 1,271 lemmas (words) with $\approx 1.94[\pm 1.7]$ senses per word

Evaluation dataset and metric

Dataset: SemCor-WSI

- 15,308 examples for 1,271 lemmas (words) with $\approx 1.94[\pm 1.7]$ senses per word

Example

Lemma: *table*

Sense 1: furniture

1. It seemed to her that the gloom drew itself in at the **table**'s empty chairs.
2. He put the shells on the **table**.

Sense 2: set of data

1. Data on such comparisons are given in **Table** 1.
2. These deductions may not be claimed if you elect to use the tax **Table**.

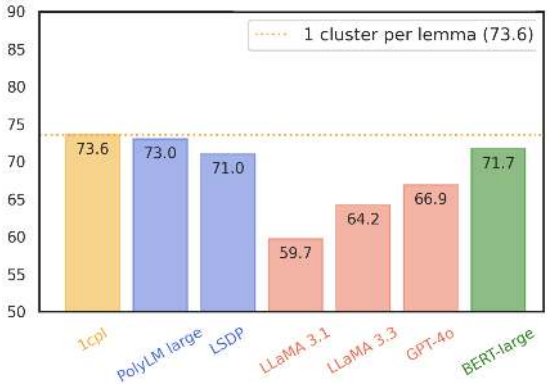
Dataset and metric

Dataset: SemCor-WSI

- 15,308 examples for 1,271 lemmas (words) with $\approx 1.94[\pm 1.7]$ senses per word

Metric: F-B^3, the harmonic mean of instance-level precision and recall, measuring same-class agreement within clusters and same-cluster agreement within classes.

WSI results



→ Predicting the same sense for all words is the best strategy
→ LLMs perform worse than previous generation models

Can we improve clustering-based approach?

Can we improve clustering-based approach?

Yes, using Wiktionary!

Can we improve clustering-based approach?

Yes, using Wiktionary!

Verb [edit]

enter (*third-person singular simple present* **enters**, *present participle* **entering**, *simple past and past participle* **entered**)

1. (*intransitive*) To go or come into an enclosed or partially enclosed space. [quotations ▼]

   *You should knock before you **enter**, unless you want to see me naked.*

2. (*transitive*) To cause to go (into), or to be received (into); to put in; to insert; to cause to be admitted.

   *to **enter** a knife into a piece of wood;*
   *to **enter** a boy at college, a horse for a race, etc.*

3. (*figurative*) To go or come into (a state or profession). [quotations ▼]

   *My twelve-year-old son will be **entering** his teens next year.*
   *She had planned to **enter** the legal profession.*

Wiktionary information

Wiktionary examples:

- BERT fine-tuning with contrastive loss (BERT-Wikt)
  - Examples of the same sense are brought closer in the embeddings space
- Must-link constraints in the clustering algortihm
  - Examples of the same sense from Wiktionary are forced to stay in the same cluster

Wiktionary number of senses:

- Number of clusters for the clustering algortihm

Wiktionary information

Wiktionary examples:

- BERT fine-tuning with contrastive loss (BERT-Wikt)
  - Examples of the same sense are brought closer in the embeddings space
- Must-link constraints in the clustering algortihm
  - Examples of the same sense from Wiktionary are forced to stay in the same cluster

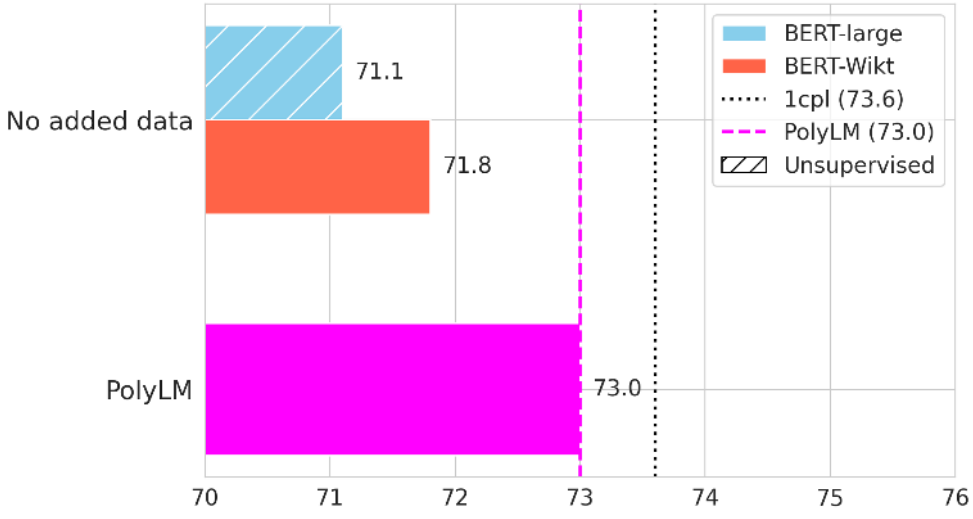
Wiktionary number of senses:

- Number of clusters for the clustering algortihm

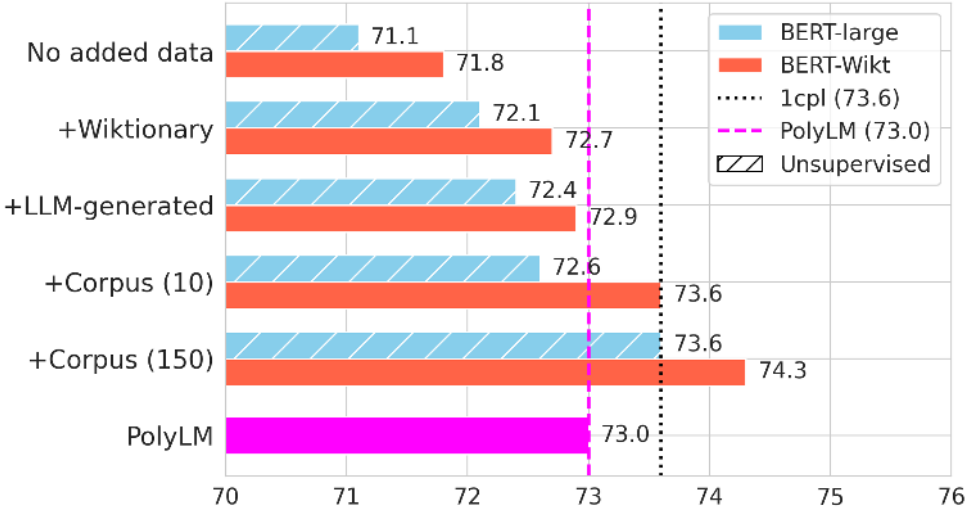
\+ **Bonus:** Adding unlabeled examples:

- From an existing corpus
- LLM-generated
- Unlabeled from Wiktionary

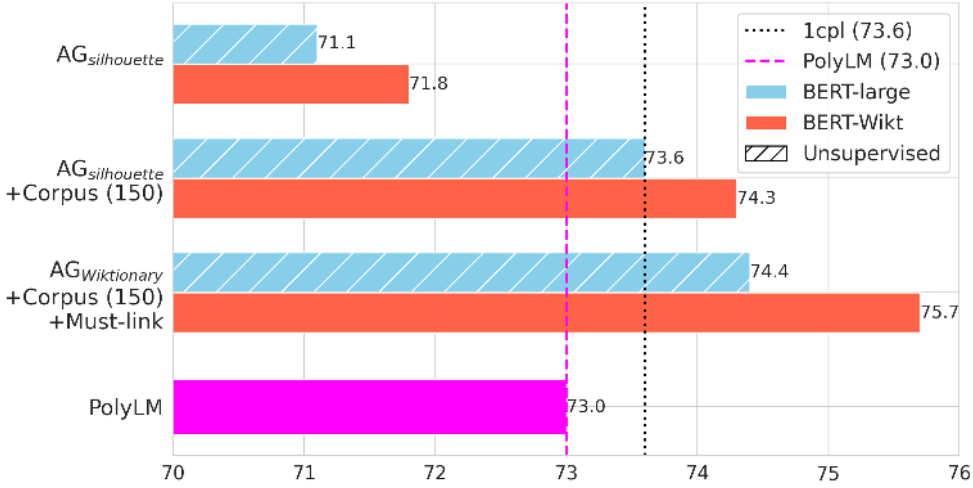
SemCor-WSI results: Baseline



SemCor-WSI results: Adding unlabeled examples



SemCor-WSI results: Must-Link constraints



Conclusion

- LLMs still struggle to understand word senses
- Clustering-based systems achieve better results than LLMs
  - They establish a new state-of-the-art result when provided with Wiktionary-based supervision and additional unlabeled examples
  - This approach can be applied to any language with a sufficiently large Wiktionary

General Knowledge Evaluation

Quiz time (1st question)

Question: "Secouez-moi, secouez-moi" was for a long time the slogan of which brand?

Possible answers:

- Pepsi
- Schweppes
- Coca-Cola
- Orangina

Quiz time (1st question)

Question: "Secouez-moi, secouez-moi" was for a long time the slogan of which brand?

Possible answers:

- Pepsi
- Schweppes
- Coca-Cola
- **Orangina**

Quiz time (2nd question)

Question: What is the diameter of the huge golden sphere located in Auroville?

Possible answers:

- 96 m
- 36 m
- 66 m
- 126 m

Quiz time (2nd question)

Question: What is the diameter of the huge golden sphere located in Auroville?

Possible answers:

- 96 m
- **36 m**
- 66 m
- 126 m

Quiz time (final question)

Among 30 LLMs (7-70B parameters) prompted with both questions, how many did answer correctly to the first question?

Quiz time (final question)

Among 30 LLMs (7-70B parameters) prompted with both questions, how many did answer correctly to the first question?

2!

Quiz time (final question)

Among 30 LLMs (7-70B parameters) prompted with both questions, how many did answer correctly to the first question?

2!

And for the second one?

Quiz time (final question)

Among 30 LLMs (7-70B parameters) prompted with both questions, how many did answer correctly to the first question?

2!

And for the second one?

30!

New evaluation dataset: TriviaRoomQA

- Source: quizzes freely available on the internet created for quiz rooms
- 3,300 parallel questions in 6 languages (19,800 in total) on 110 different topics (history, geography, food, entertainment, music, news, sport, people, tech, etc.)
- 8,640 French-only questions on 288 topics (including the subset above)
- Each question has 4 possible answers
- Questions are divided into 21 categories, 3 levels of difficulty, 10 time periods, 9 continents
- Metric: accuracy over conditional log-likelihood of each answer

Models tested

30 LLMs in total (7-70B), today we'll see only the following:

- **European models:** Salamandra-7B, Apertus-8B, Gaperon-8B, EuroLLM-9B, Minerva-7B, LLaMmlein 7B
- **North American models:** Aya-8B, Granite-8B, Olmo-3-7B
- **Asian models:** ALLaM-7B, Qwen-3.5-9B

Results groups

Languages: English, French, German, Dutch, Italian, Spanish

Results groups

Languages: English, French, German, Dutch, Italian, Spanish

Difficulty level: Beginner, Confirmed, Expert

Results groups

Languages: English, French, German, Dutch, Italian, Spanish

Difficulty level: Beginner, Confirmed, Expert

Categories: Nature, Colors, Music, Food And Drinks, Movie, Personalities, News, Sport, Language, Cities And Countries, Health And Human, Numbers, Geographical Places, History

Results groups

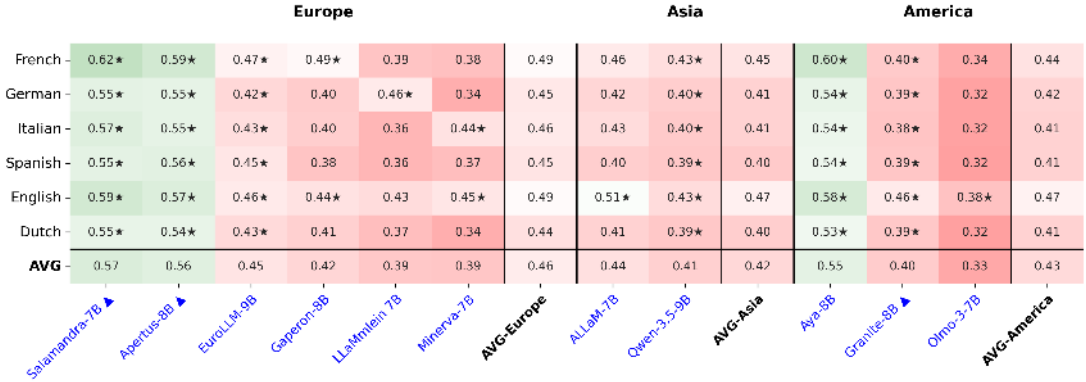
Languages: English, French, German, Dutch, Italian, Spanish

Difficulty level: Beginner, Confirmed, Expert

Categories: Nature, Colors, Music, Food And Drinks, Movie, Personalities, News, Sport, Language, Cities And Countries, Health And Human, Numbers, Geographical Places, History

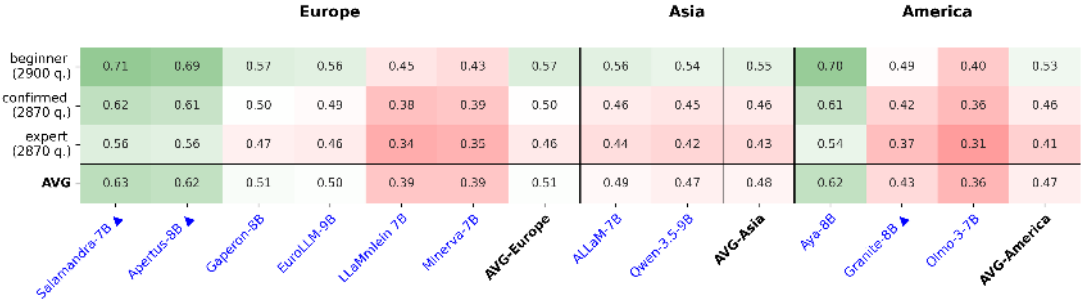
Time periods: Timeless, Historical, 1950s–1960s, 1970s–1980s, 1990s–2000s, 2010s–2020s

Results across languages



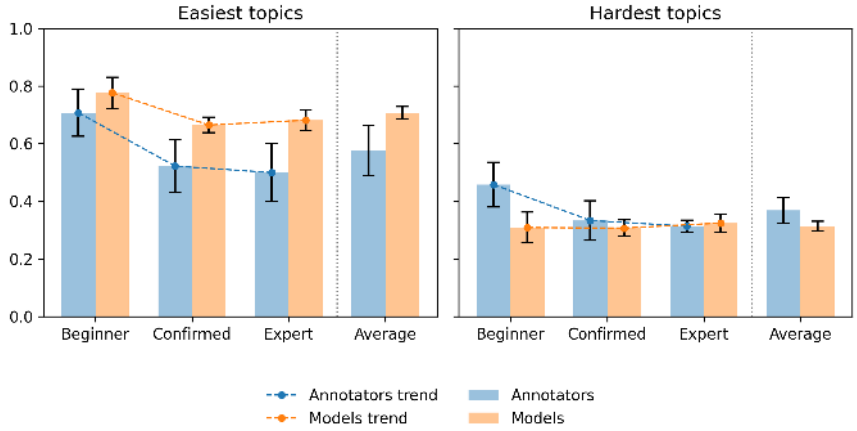
→ The same questions are not answered in the same way in different languages

Results across difficulty levels

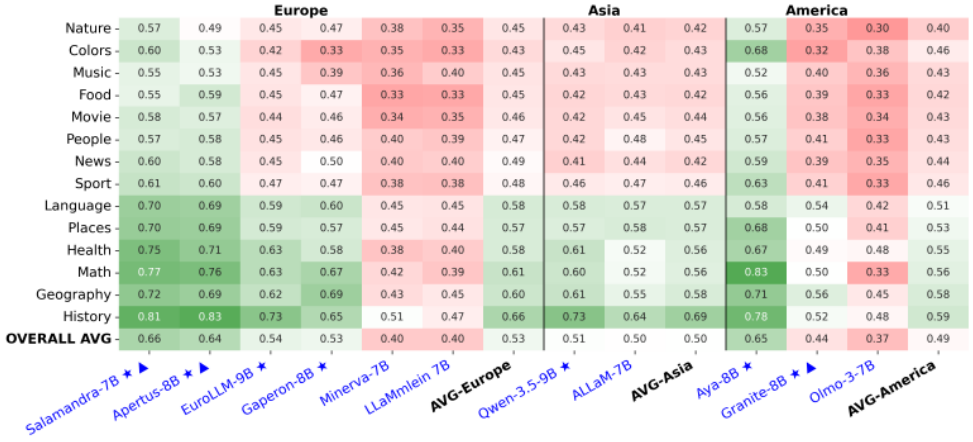


→ As expected, models' scores decrease on more difficult questions. But!

Difficulty levels: humans and LLMs patterns on unfamiliar topics

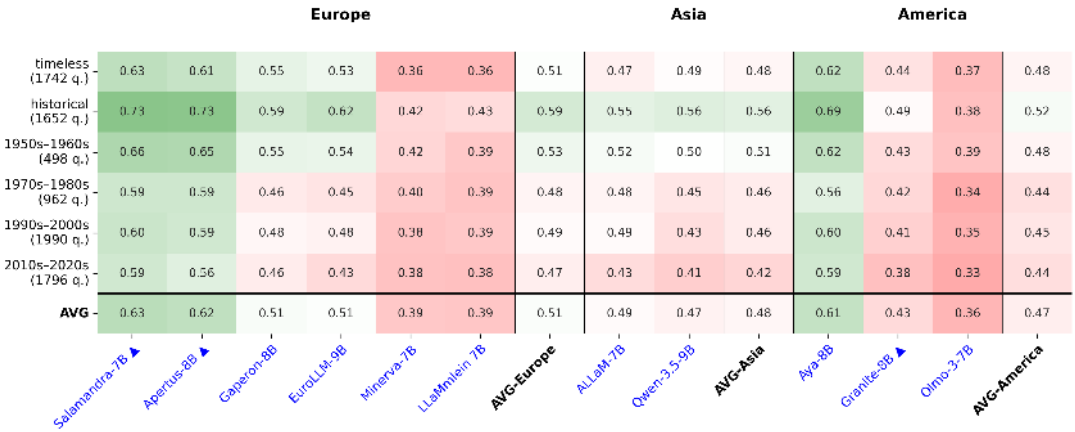


Results across categories



→ Encyclopedic knowledge is easier for models compared to everyday knowledge questions

Results across time periods



→ Historical and timeless questions are easier for models

Bonus: improving models performance with search tool access

Two models: Qwen3.5-9B and Llama-3.1-8B

Dataset: French part of the TriviaRoomQA

Name	Original model	Model with search tool
Qwen3.5-9B	0.47	**0.79**
Llama-3.1-8B	0.51	**0.63**

Conclusion

- Trivia is challenging for open-weight LLMs
- LLMs perform well on encyclopedic categories but are close to random on everyday topics
- LLMs provide inconsistent answers to the same questions across languages
- Unlike humans, LLMs perform poorly across all difficulty levels when dealing with unfamiliar topics
- Historical and timeless questions are easier than recent ones

Conclusions and future work

General conclusion and Future work

Conclusions:

- When prompted to cluster examples according to the senses of a target word, LLMs perform worse than simpler clustering-based models
- LLMs excel at questions involving encyclopedic and historical knowledge but struggle with everyday general-knowledge questions, such as those about food or celebrities

Future work:

- Evaluation of LLMs in other specialized domains: finance, culture, education, healthcare, public administration

References I

- Amrami, Asaf and Yoav Goldberg (2019). *Towards better substitution-based word sense induction*. arXiv: 1905.12598 [cs.CL]. URL: https://arxiv.org/abs/1905.12598.
- Ansell, Alan, Felipe Bravo-Marquez, and Bernhard Pfahringer (Apr. 2021). "PolyLM: Learning about Polysemy through Language Modeling". In: vol. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. arXiv:2101.10448 [cs] version: 1. Association for Computational Linguistics. URL: https://aclanthology.org/2021.eacl-main.45/ (visited on 10/20/2022).