MSR193

Evaluating the Performance of a Large Language Model (LLM) Compared to Humans in a Complex Categorisation Task

Edema C[1], Martin A[1], Martin C[1], Bertuzzi A[1], King E[1], Wesson F[1], Witkowski M[1]
[1] Crystallise, Stanford-le-Hope, UK

Background

Manually indexing abstracts into multiple fields can be time-consuming and prone to errors. LLMs have shown remarkable speed and accuracy at analysing texts.

We have previously shown that an LLM was accurate at categorising abstracts according to disease area studied. Hence, our current aim is to determine its accuracy in indexing a more complex field that requires more subjective interpretation.

Methods

- We conducted a literature search and retrieved 500 abstracts assessing the ***impact of interventions to delay ageing***.
- Using an ***online evidence mapper tool*** (www.evidencemapper.co.uk), the abstracts were categorised independently by human researchers and the LLM to the ***12 hallmarks of ageing***.
- A ***Geroscience expert*** generated a list of keywords relevant to each ageing hallmark which was used to train the LLM.
- The time taken for each approach was recorded. A gold standard categorisation for comparison was created independently.

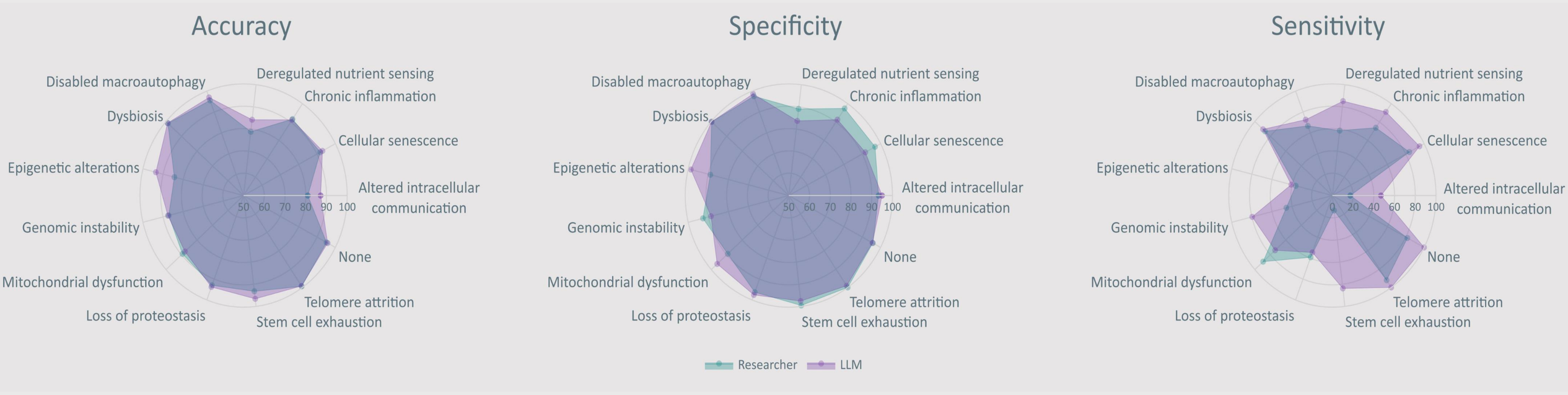
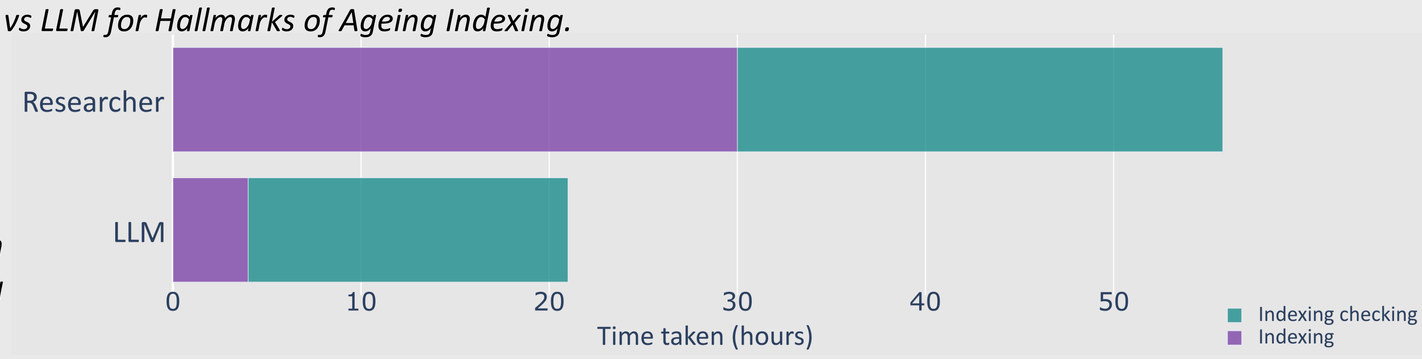


Figure 1: Comparison of the Sensitivity, Specificity, and Accuracy of Researchers vs LLM for Hallmarks of Ageing Indexing.

Figure 2: Comparison of the Time Taken for Hallmarks of Ageing Indexing



Results

- Of the 500 abstracts, 478 publications reported at least one hallmark of ageing.
- The average ***sensitivity, specificity and accuracy*** of the LLM in grouping the abstracts by hallmarks of ageing were ***77.9%, 94.9% and 92.8%*** respectively. In comparison, human researchers recorded a mean sensitivity, specificity and accuracy of ***61.9%, 95.2% and 90.7%*** respectively.
- The initial indexing by the LLM was completed in about ***one-seventh of the time taken by human researchers*** (4 hours versus 30 hours), while checking of the LLM's indexing and the researchers' indexing took 17 hours and 25.8 hours respectively.

Conclusions

- The human-trained LLM performed better and faster than human researchers at indexing abstracts to more complicated fields.
- This underscores the importance of leveraging artificial intelligence to achieve consistency in accuracy when undertaking complex indexing tasks.
- Further research is required to ascertain the cost-effectiveness of utilising LLMs for categorising abstracts.

Email:
contact@crystallise.com

Website:
www.crystallise.com

Crystallise

LinkedIn (Crystallise Ltd)