Complexity Science★Hub

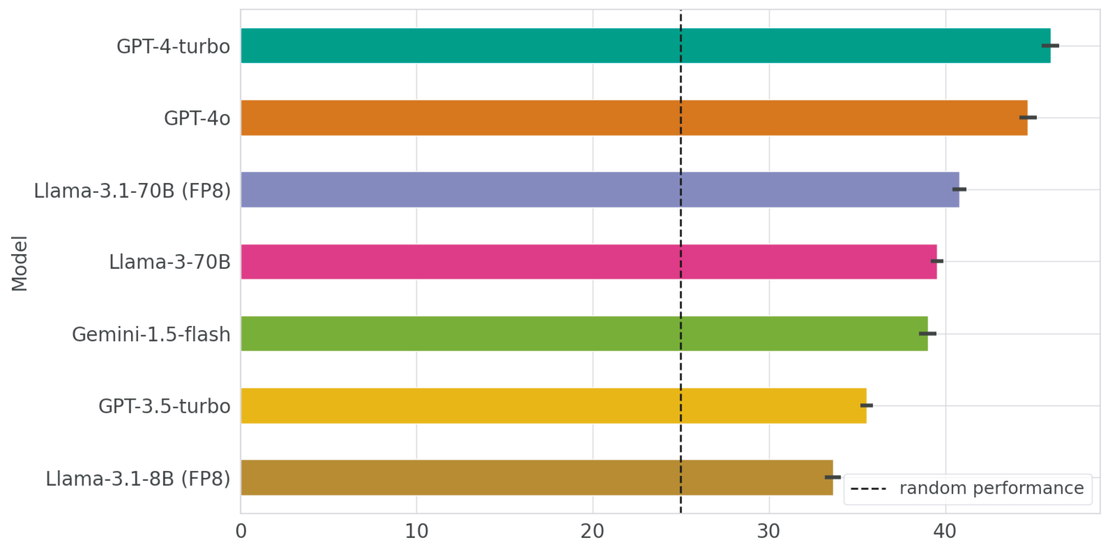
State-of-the-art LLMs still lack expert level historical knowledge

Jakob Hauser[1]†, Daniel Kondor[1], Jenny Reddish[1], Majid Benam[1], Enrico Cioni[2], Federica Villa[2], James S. Bennett[3], Daniel Hoyer[4], Pieter Francois[2], Peter Turchin[1], R. Maria del Rio-Chanona[15]†

[1] Complexity Science Hub, Vienna, Austria
[2] University of Oxford, Oxford, UK
[3] University of Washington, Seattle, WA, USA
[4] George Brown College, Toronto, Canada
[5] Department of Computer Science, University College London, London, UK

† hauser@csh.ac.at; m.delriochanona@ucl.ac.uk

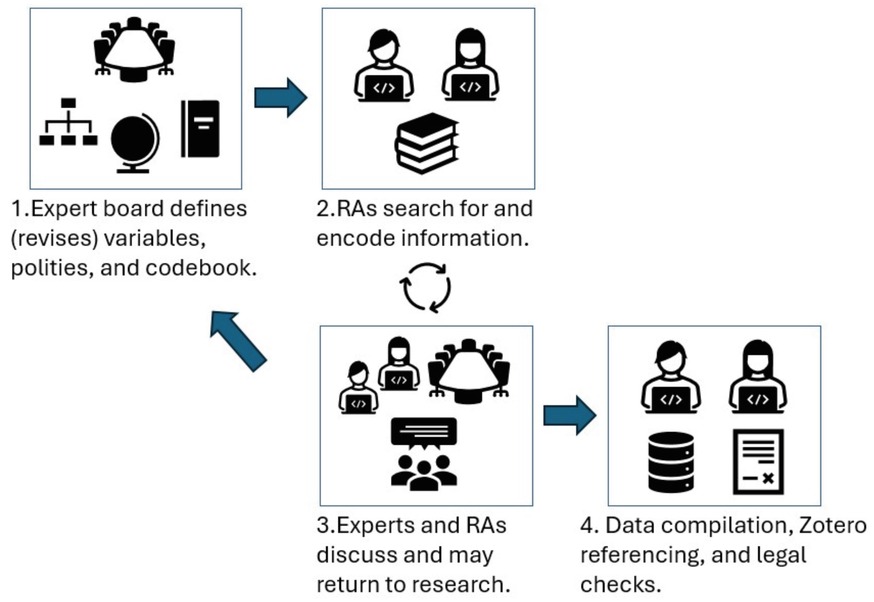
BALANCED ACCURACIES FOR 4-CHOICE TASK



We present a dataset and benchmark building upon the Seshat Global History Databank, a structured representation of human historical knowledge, containing 36,000 data points across 600 historical societies and paired with references to over 600 scholarly works, covering every major world region from the Neolithic period to the Industrial Revolution.

We benchmark the historical knowledge of 7 closed and open weight LLMs of various sizes. We find that LLMs demonstrate balanced accuracy ranging from 33.6% (Llama 3.1-8B FP8) to 46.0% (GPT-4-Turbo) in a four-choice format, outperforming random guessing (25%), but falling short of expert comprehension. LLMs perform better on earlier historical periods, with accuracy decreasing for more recent times. Regionally, performance is more even but still better for the Americas and lowest in Sub-Saharan Africa for the more advanced models. Our benchmark suggests that while LLMs possess some expert-level historical knowledge, there is considerable room for improvement.

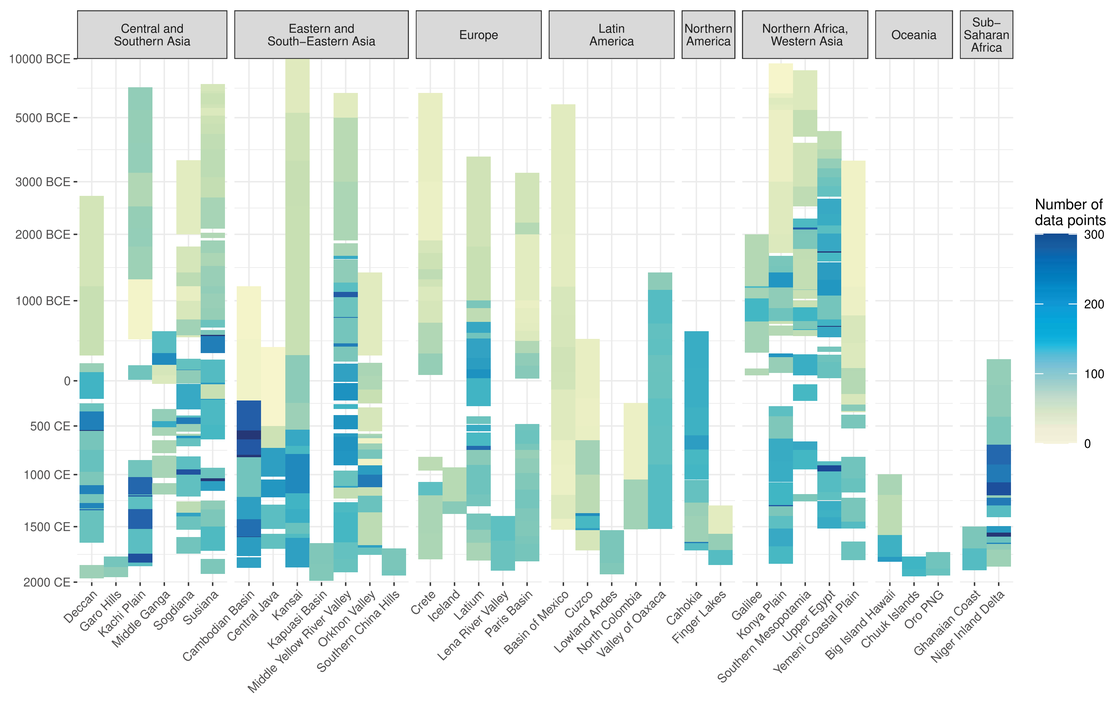
THE SESHAT DATA COLLECTION PROCESS



EXAMPLE OF A SESHAT DATAPOINT

Polity	Middle Kingdom Egypt
Variable	Settlements in a Defensive Position
Value	Present
Description	Middle Kingdom fortresses "were ...
Reference	Van de Mieroop 2011, 113

TEMPORAL SPREAD OF DATA



FEW-SHOT CONTEXT EXAMPLES WITH EXEMPLARY REASONING

Question:
The characteristic 'elite status is hereditary' is categorized under 'Status'. Was it present, inferred present, inferred absent, or absent for the polity called 'Parthian Empire II', during the time frame from 41 CE to 226 CE?
Reasoning and evidence:
Elite families such as the Surens and Karens had the greatest influence and probably held top posts such as "satrap of satraps" and were regular satraps.
Answer:
Present
...
Question:
The characteristic 'Chattel slavery' is categorized under ' Proportion of population enserfed '. Was it present, inferred present, inferred absent, or absent for the polity called 'Kingdom of Hawaii - Kamehameha Period', during the time frame from 1778 CE to 1819 CE?
Reasoning and evidence:
In late precontact and early contact-era Hawai'i: 'The papa kauwā, at the bottom of the social scale, are sometimes translated in Western literature as "slaves," but a better term is probably "outcast".
Answer:
Absent

BENCHMARK QUESTION

Question:
The characteristic 'elite status is hereditary' is categorized under 'Status'. Was it present, inferred present, inferred absent, or absent for the polity called 'East Roman Empire', during the time frame from 395 CE to 631 CE?
Reasoning and evidence:
<model answer>

poster edited by ella weger :---)