Plausibly Historical?

LLMs as Readers and Writers of History

The Plausible & the Implausible

- Understanding history – and trusting it – matter as much as ever. What impact are LLMs having on that process?

- Part 1: Using LLMs to read history
  - Can we use LLMs to structure and 'read' historical data?
  - What are the possibilities – and the drawbacks?
- Part 2: The LLM as historian
  - LLMs are already in use to present history to the public
  - Increasingly a main workhorse of pop-history presentation
  - How does the character of LLMs affect this? What stories are they telling?

Part 1: The LLM as Historical Reader

- History is often formed of nuanced, individual statements in large volume
- Hard to summarise or statistically analyse conventionally
- And thus hard to find contradictions, problems, and structural patterns
- LLMs offer new capabilities for text processing
- Some sorts of text processing in history are difficult at scale

- Are LLMs a solution that can build and reflect trust? Or a dangerous dead-end that will skew our thinking about the past?

The Problem

- It's useful to have historical information as structured data
- Why?
  - First, searchability
  - Second, comparison and visualisation of networks and data points
  - Third, using the structuring process can challenge and tell us about our own assumptions
- Nb what's not here: statistical analysis and averages
- Unstructured text -> structured data done by human reading
- Slow, time consuming: is there another way?

“Specificities” of history

- Historical knowledge is different to much STEM knowledge
- Data systems are often built with scientific data in mind
  - Key element: scientists usually trust one another not to lie in their papers!
- For historians, who said it and where they claim to know from matter
- Which claims do we trust? Who agrees with who?
- What are the relationships between that is being said and the truth?

- These are not true “specificities”: this covers *most human knowledge*

Data structures

- Triples (Subject, Predicate, Object)
  - Standard data framework for connected entities
  - Lacks provenance, authority
- Direct authority
  - Single assertion of truth-claims, authorship for whole dataset
  - Problematic if multiple authorship, represents just one argument
- STAR (Structured Assertion Record)
  - Adds provenance and authority
  - Permits argumentation modelling “X says [TRIPLE] based on Y” where Y can be another STAR

Why read with LLMs/NLP?

- We simply do not have that many historians!
- Populating high-complexity data structures like STAR is a slow task

- LLMs can construct plausible/usually-correct answers to entity recognition and, perhaps, authority mining questions
- Enforcing data structures increasingly possible out of the box
- Allows the construction of efficient networks of argumentation that can be used to find and test contradictions and issues with historical information

What might the process look like?

- Ascertain entities, predicates, references
  - Not clear if LLMs outperform other NLP methods on some of these
- Create attribution of resulting triple
  - Problem of authority mining
  - Can leverage references in secondary texts
- Potential extension: identifying sourcing

- Potential extension: identifying agreement
  - People can make assertions about assertions
  - Which assertions being referenced does an author agree with?

What are the problems?

- Human writing is often ambiguous in sourcing
  - E.g. a sentence with three statements and a reference at the end: does it apply to all three statements, just the last, just the one that most needed referencing?
- Citations can be citing for fact – but also citing for disputes
  - I might well cite a text to specifically note that it was wrong
  - I'm not necessarily even saying that I think that author believed what they wrote!
- Outside "history facts" further problems arise
  - "Puppies are nice" -> assertion we agree with, but isn't a provable truth
- Balance between cross-referencing and novelty
  - More robust if we can check against existing data, but overfit struggles with new people or ideas

Is LLM performance good enough?

- It depends on your bar line for "good enough".
  - Lack of agreed benchmarking for authority mining
  - And unclear how well benchmarks transfer between subtly varied tasks
  - Ironically, may need more humans doing this work for benchmarking
  - Still unclear what the potential "maximum performance" is
  - But the question is rarely maximum, but practical likely, performance
- Data *checking* is still a time consuming problem
- The method comes under its own pressures: who is saying this?
  - And how do we communicate when an LLM is? What value does that have?

What do we lose methodologically?

- Direct contact with source material
- Digital or not, close reading cannot be done without a reader
  - There are some things we can't realise through summaries
  - Harder for the LLM to "notice" a connection not represented in training data
- The LLM cannot answer the questions we do not ask it
  - We can enforce a data structure, but hard to then flag up when the data structure itself has problems
  - Because data structures represent our thought processes, testing and refining them can be an important element of data-driven history

Possibilities and pitfalls

- Solving authority mining is key to using LLMs to process history data
- Many other applications too (law, journalism, etc)
- Structured data with authorities attached could help us use LLMs to think about perspective and truth at data scales: combatting information overload
- UI and presentation are absolutely key: we have to communicate what people are seeing, and who is behind that information

- Without a grasp on authority and truth-values, LLMs and data can struggle with key historical issues of truth, authority and citation
- And this is already a problem in practice …

Part 2: The LLM as History Writer

- Public engagement with history often via social media
- Huge surge in auto-production of history & culture content
- Often “low-grade slop” end of LLM use – but people see it!
- Especially common on larger social media sites
- Also websites designed for SEO, sometimes for data gathering/advertising

- So what histories are LLMs writing? And why?

The Problem

- LLMs make production of rapid text & images easy
- Those texts and images can produce historical-cultural messaging
- Subject to a) the model and b) the prompts used
- Key vector for inbuilt issues with training data being replicated

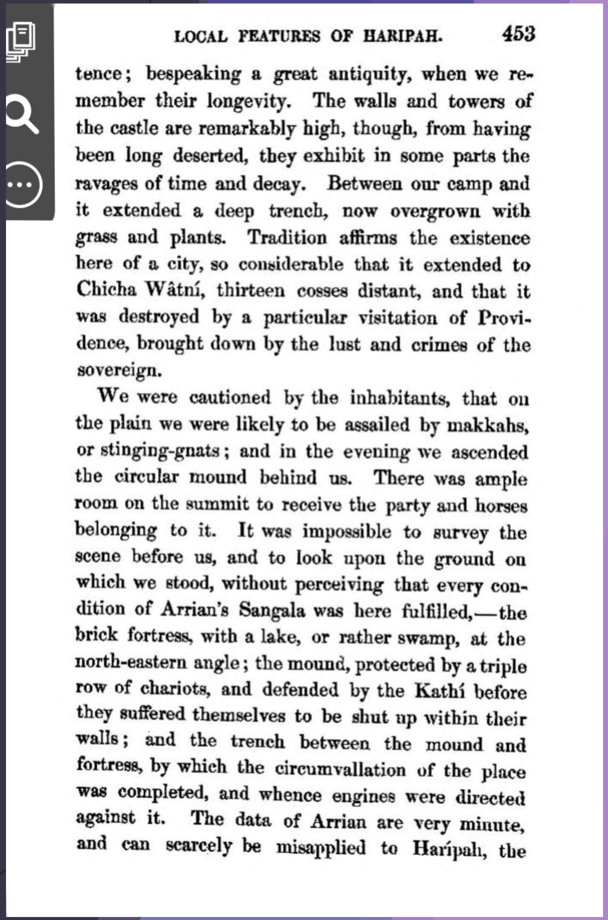
- Not just a question of what LLMs *can* do: what *are* they doing?

What does this look like?

- Often short image + graphic formats
- Longer ‘explainer posts’ just from generated text
- Increasing use of short video as LLM capabilities have grown
- Quite possible also to generate book-scale content

- Usually unclear where the AI/human boundary lies
- Which model or prompts used are not given
- High confidence and authoritative styles for text
- Photorealistic or animation styles especially common

What does this look like?



LOCAL FEATURES OF HARIPAH. 453

tence; bespeaking a great antiquity, when we remember their longevity. The walls and towers of the castle are remarkably high, though, from having been long deserted, they exhibit in some parts the ravages of time and decay. Between our camp and it extended a deep trench, now overgrown with grass and plants. Tradition affirms the existence here of a city, so considerable that it extended to Chicha Wâtní, thirteen cosses distant, and that it was destroyed by a particular visitation of Providence, brought down by the lust and crimes of the sovereign.

We were cautioned by the inhabitants, that on the plain we were likely to be assailed by makkahs, or stinging-gnats; and in the evening we ascended the circular mound behind us. There was ample room on the summit to receive the party and horses belonging to it. It was impossible to survey the scene before us, and to look upon the ground on which we stood, without perceiving that every condition of Arrian's Sangala was here fulfilled,—the brick fortress, with a lake, or rather swamp, at the north-eastern angle; the mound, protected by a triple row of chariots, and defended by the Kathí before they suffered themselves to be shut up within their walls; and the trench between the mound and fortress, by which the circumvallation of the place was completed, and whence engines were directed against it. The data of Arrian are very minute, and can scarcely be misapplied to Harípah, the

What does this look like?

- Obvious inaccuracies
  - Minarets as factory towers?
  - Yep, that’s an American eagle
- But deeper issues too
  - Dehistoricised modern symbols

Why is this being done?

- For money: some pages seek to directly sell related things
- For money: or to build an audience they can then sell
- Politics: some pages push views of countries and politicians
- Politics: some pages seek to push social politics issues
- Advertising: some pages are directly selling products/travel
- Curiosity: some users think generated things must be true
- Curiosity: users who don’t spot or don’t care about the AI will then organically keep these things moving

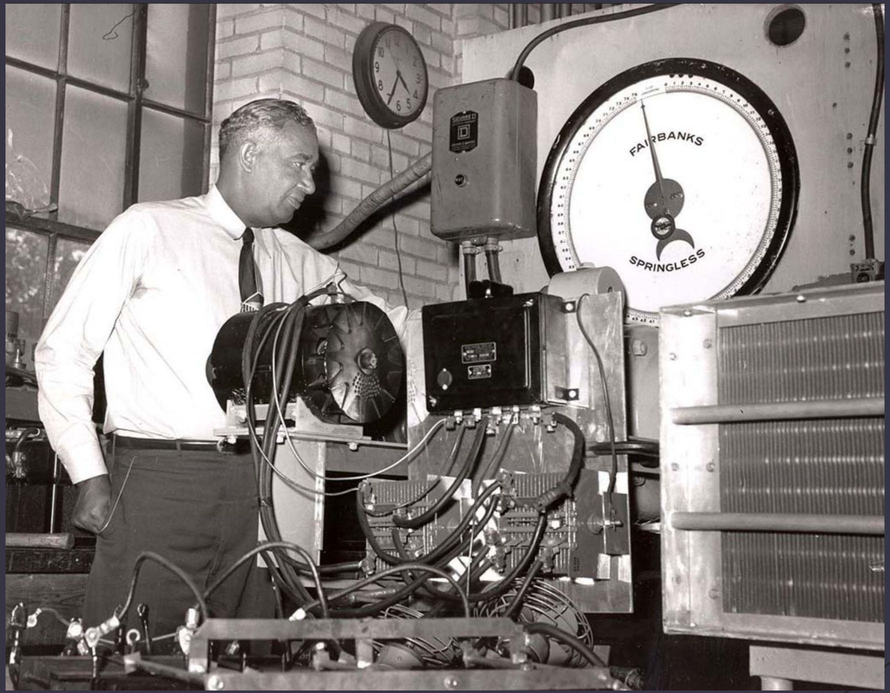
Averages as biases

- Tendency for LLMs to produce modal averaging effect
- An average is not ‘unbiased’, it is just an average bias
- If most people whose work the AI is trained on are wrong, the AI is also likely to be wrong
- And sometimes most writers are very wrong!
- Or are right within their own concept frameworks
  - Americans really like bald eagles, but other eagles also exist …
  - Or were accurate but now outdated
  - E.g. presenting a situation trying to be accurate but lacking info

Assessing images

- LLM averaging effect likely to pick up on tendencies in pop media
  - See e.g. Frederick McKinley Jones
- Not just iconography: also in e.g. colour
- We can use an RGB colour space as a vector to examine this
- Initial tests suggest that LLM images of premodern historical topics may tend to exceptionally low colour variance – dark and muted tones, variance closer to a field of wheat than a photo of a landscape
- Comparisons on variation metric:
  - Rainbow 53%, Bright landscape photo 48%, dull landscape 46%, grass with blue flowers 35%, AI Armenian Monarchs 19-25%, wheat 19%

Frederick McKinley Jones: LLM vs Photo



Whose story do LLMs tell?

- Not just a problem of untruths but of which truths
  - “Patriotic” maps and narratives and “golden ages” often very popular
  - Dramatic stories and well known figures help drive views
  - Also spark arguments, key for social media engagement
  - But reinforce limited, statist, “great man”, nationalist/exclusionary visions
- And what gets filled into the blanks of history
  - Especially an issue in visual representations
  - In Frederick McKinley Jones’ case we literally lose the face of a historical man
- History doesn’t always work best in soundbites
- LLM averaging tendencies ‘mode average out’ even *median* normal stories

Why is this a problem?

- ‘Tropey’ representations are inherent to generative LLMs
- LLMs learn common nationalist and populist history tropes
- And accept/reinforce older historiographies
- Dealing in (often dubiously sourced) facts over understanding

- E.g. the “Spartan Mirage” (or similar “Fremen Mirage”)
  - Idea of the Spartans as brutalised but tough manly super-soldiers who won
- Generally false: not how Spartans won wars, drove Sparta to collapse
- Brought up today in ‘manosphere’ and far-right movements
- Replicated and pushed forward by LLMs on social media

History in Volume

- LLMs can produce more historical ‘content’ than us.
- Historians rarely have time, skills, or resources to compete
- Visual & engagement-focused representations especially rare in academic work, common in popular culture

- A high proportion of LLM content is misleading at best
- LLM companies have effectively declared war on the public understanding of history
  - (And they’re probably not even aware they’ve done so)

What can we do?

- Improve public understandings of LLMs
- Focus on the platforms and infrastructure we use: holistic research view
- Rethink academic engagement: the age of ‘our public presence is four tweets and a blogpost’ needs to end
- Be clear:
  - about what we are using LLMs for
  - about why we are using them
  - about the times when we shouldn’t use them
  - that they are tools with specific uses

Conclusions

- For Iranian studies, significant and immediate problems
  - LLM content will reach people long before ours does
  - War in Iran issues already plagued by faked content
  - Cannot be separated from ‘higher’ academic work: a public fed too many ‘clash of civilisation’ and ‘spartiate’ narratives will not see the purpose of research
- Key overarching question is trust in information
  - Can we build LLM pipelines that retain and help people assess trust?
  - Can we get scaling benefits without de-skilling or losing close reading contact?
  - What are the impacts of LLMs automating simple generic historical narratives?
  - How do we push back against LLMs tendencies to misinform?

Thanks for Listening

Dr. James Baillie

Institute for Iranian Studies

james.baillie@oeaw.ac.at

@jubalbarca@scholar.social

@jubalbarca@bsky.app