Running Reality

https://www.runningreality.org

Extracting Structured Data from Historical Narratives Using the Running Reality Application in Combination with a Large Language Model

HENNING, Garth[1]

[1]Running Reality Organization, Garth Henning

ghenning@runningreality.org

https:// www.runningreality.org

Abstract. Digital history tools use structured data to create models of historical environments, but a very large fraction of historical data is in narrative format. Building a large set of structured data requires identifying individual factoids from within historical narratives. Recent advances in Artificial Intelligence and Machine Learning (AI/ML) have led to innovative neural networks known as the Large Language Models (LLMs) that can follow a train of thought in written work and then answer questions about that work. The Running Reality digital history desktop application has been upgraded with an experimental feature to interface with LLMs to import data from narrative text. Running Reality breaks up the text into single-topic sections, provides the section to the LLM, then asks the LLM a predefined set of questions. Running Reality has predefined sets of questions for text whose subject may be a city or a person, to determine if the text contains basic data such as founding or birth dates, alternative names, as well as locations over time. The OpenAI ChatGPT version 3.5 LLM is able to work with text within a 4096 token (or approximately 3000 word) look-back attention buffer, so Running Reality tries to keep section text to within this limit. The results of the experimental feature show that a combination of Running Reality and an LLM promises to be able to build large structured historical datasets.

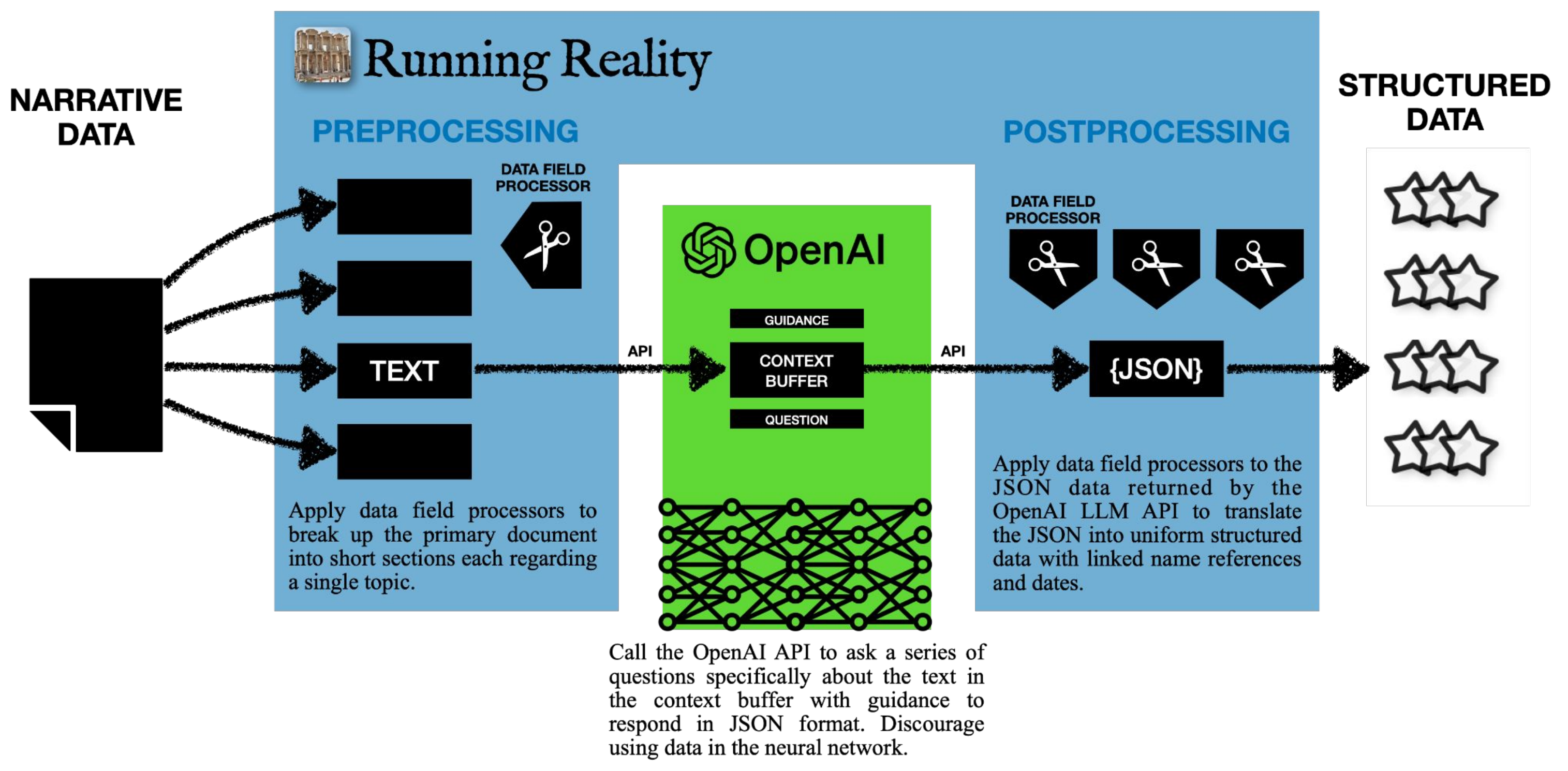
Key words. Digital history tooling, machine learning, structured data, data extraction, world history model;

Hypothesis: New Large Language Model (LLM) AI services can extract structured data from historical narrative with an accuracy comparable to a human and at a lower cost.

For a human to extract structured data in a uniform format takes time and tooling beyond just reading the text. The accuracy of humans depends on skill level. Crowdsourcing approaches in other fields have relied on large numbers of volunteers to cross-check one another, extensive support tooling, and expert review of results. Even higher-skilled paid humans would require tooling to produce uniform results, i.e. validation and linking of dates, names, event wording, and locations.

Most historical data is in narrative form and existing structured data sets have been built at great cost and, as a consequence, can carry usage or license restrictions.

Running Reality adapted its existing data source processor that can ingest, transform, and reformat structured historical data. The Running Reality app calls the LLM-as-a-service known as OpenAI GPT via its Application Programming Interface (API). The app sends blocks of narrative wrapped with guidance instructions 1) to only use the text provided and 2) to produce JavaScript Object Notation (JSON) output and a series of questions about whether the text references historical events, such as whether a city experienced an earthquake. This capability is experimental, but is currently available to all users of the app. Running Reality is characterizing the performance of this experimental approach by assessing against narrative data sources of value to Running Reality.

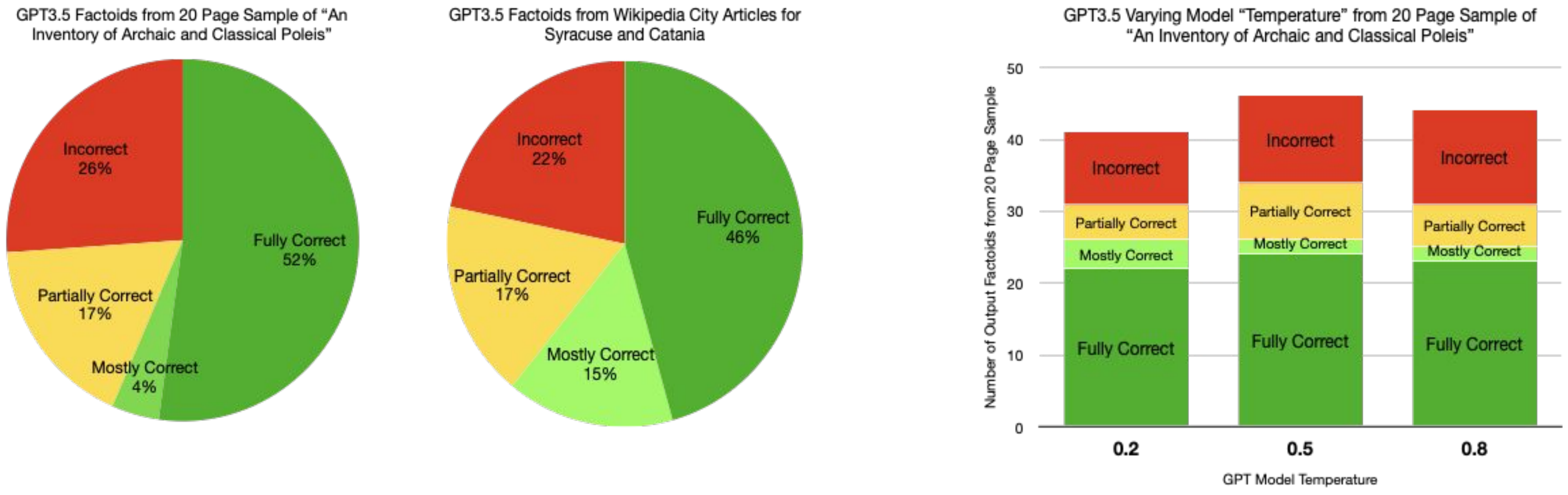


Apply data field processors to break up the primary document into short sections each regarding a single topic.

Call the OpenAI API to ask a series of questions specifically about the text in the context buffer with guidance to respond in JSON format. Discourage using data in the neural network.

Apply data field processors to the JSON data returned by the OpenAI LLM API to translate the JSON into uniform structured data with linked name references and dates.

Data Source	Pages	Questions	Tokens	Model	Factoids	Fully Correct %	Price
An Inventory of Archaic and Classical Poleis	20	8 per section	377755	GPT3.5	46	52%	$0.38USD
				GPT4	19	100%	$3.82USD
Wikipedia	2	8 per section	143099	GPT3.5	46	46%	$0.14USD
				GPT4	16	88%	$1.45USD



Fully correct factoids had all data correct and could be used in the Running Reality world history model. Mostly correct factoids had a minor error, such as a formatting error. Partially correct factoids had some data correct, such as a date or event subject or event object but had some data incorrect, such as mistaking the outcome of an event. Incorrect factoids were unusable, with no traceability to the source text.

RESULTS show promise, yet human supervision remains critical.

- The best results are for already quasi-structured data, such as latitude and longitude values embedded in text. However, this data is often already available fully structured.
- The worst results required differentiating event outcomes. It identified conflicts, conflict dates, and combatants, but struggled to differentiate seiged, sacked, and conquered.
- No significant difference in outcome for varying model "temperature," which OpenAI considered a parameter to tune GPT's creativity by increasing the variability of the output. Directing the use of only information in the context buffer and directing structured JSON output already reduced variability.
- Results from the "simpler" Wikipedia articles were not appreciable better than denser academic text. Investigation of individual errors pointed to imprecise grammar in Wikipedia articles and ambiguous event causality lowering the accuracy.
- GPT4 was 10x as expensive as GPT3.5 and produced too few factoids for effective comparison. Investigation of individual factoids showed GPT4 had a high accuracy percentage *but only by avoiding even obviously described events.* GPT4 results primarily matched the latitude and longitude data that was also identified accurately by GPT3.5

NEXT STEPS will test additional kinds data sources and improve the RR interface with GPT

- Expand from cities to people and buildings to extract locations, events, and relationships.
- Improve accuracy on a single record by:
  - Iterate exact language of queries to better distinguish events
  - List all events of a given type, not just a single event per query
  - Request a confidence level
- Assess more records
  - With higher per-record confidence, incur costs to assess larger numbers of records
  - Develop better statistics on which kinds of data have higher confidence
- Investigate ways to make better use of GPT-4 capabilities
- Iterate app user interface
  - Facilitate human supervision to make factoid checking easier
  - Link back candidate factoids to the source text
  - Flag confidence level and potential alternate text to users

Running Reality is our model of world history that plays out on a digital map freely accessible using mobile devices and desktop computers.

- Hundreds of daily visitors
- Live online since 2014
- Freely view dynamic interactive map of all world history

- The goal is to represent history from ~3000BCE to today
- Scalable fidelity, depending on the data
- Precision down to meter-level and hour-level, when warranted
- A single, unified, integrated global model

DESKTOP APP VERSION is a tool to edit, analyze, and explore historical data.

BASELINE WORLD is comprised of 1.37 million factoids and growing, modeling 322,000 historical objects and backed by over 3000 citations.

EDITING TOOLS to create new factoids with GIS-style tools (including deep OpenStreetMap and historical map collection integration) and non-geographic data tools.

DATA SOURCE integrations to layer external data or transform it from image formats (PNG, JPG, TIFF), GIS formats (GeoJSON), data formats (CSV, XLS, RDF, EpiDOC, GEDCOM, SQL), and text formats (TXT, HTML, PDF).

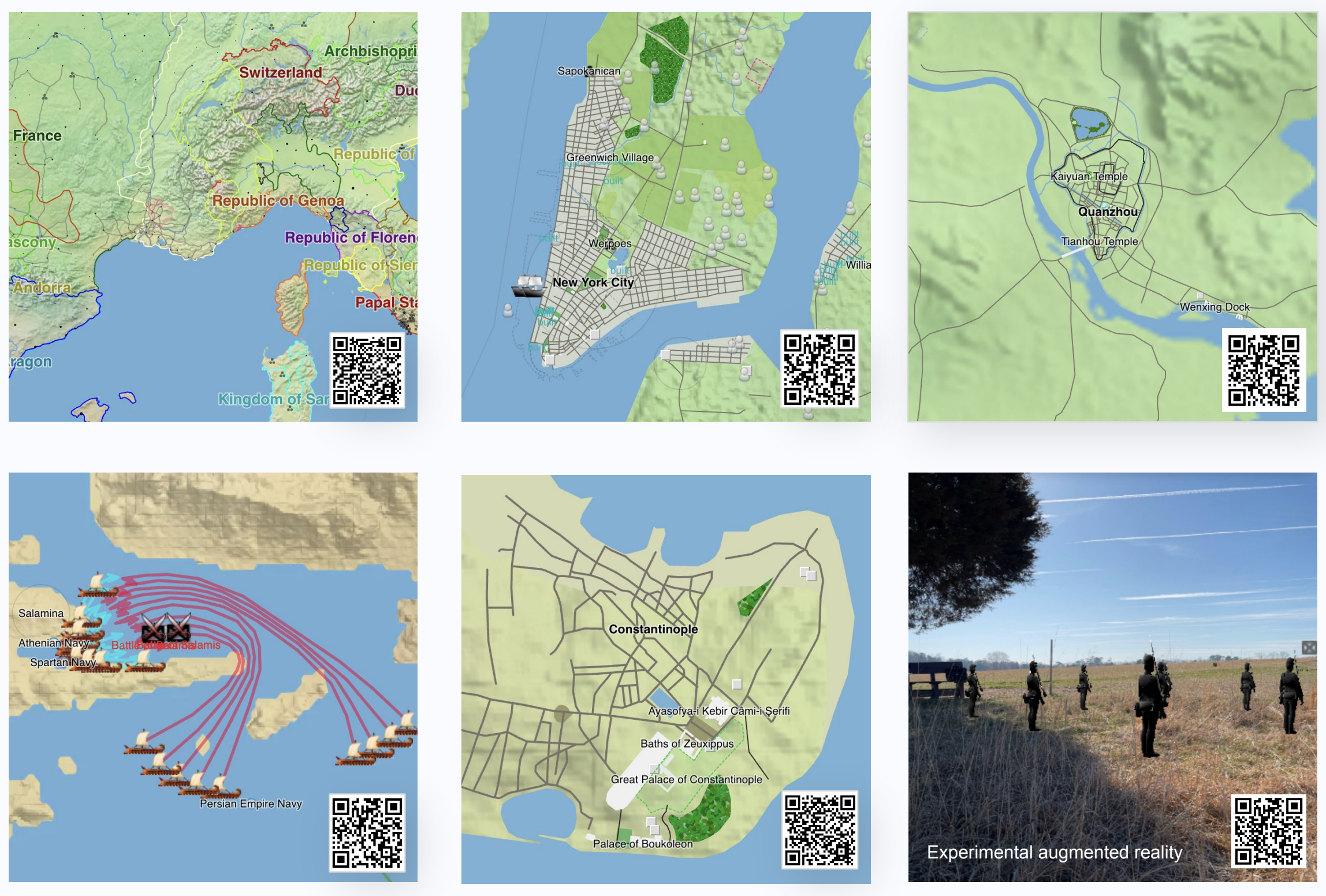
WEB VERSION is a read-only way to explore history on any desktop or mobile device

HIGHLIGHTS help users quickly navigate to interesting history topics.

PROJECTS help interested users choose ways to contribute, sorted by skill level.

LESSON PLANS are a new way for teachers to guide a class on a journey exploring history

HOSTED RESEARCH is a new story-map and file hosting service



LLM INTERFACE CONSIDERATIONS

- LLMs are known to confidently "hallucinate" responses that are comprised of probable words that fill gaps in its training data. Limiting "hallucinations" is critical.
- LLMs have memorized terabyte-scale training data as part of their neural network weighting parameters, which, while not explicitly traceable or cited by LLM vendors, includes vast historical data yet leaves gaps and inconsistencies of unknown scale.
- Direct the LLM to respond based only on the input text so that the answers are based on data in the LLM context buffer, not the neural network.
  - Narrative is broken into sections that fit within the token count (roughly the word count) of the LLM context buffer (4096 tokens for GPT3.5turbo, 16k tokens for GPT3.5turbo-1106)
  - The content of the LLM context buffer is small but can be explicitly known, while the neural network is vast but unknown.
  - Reduce the "temperature" setting of the LLM using its API to narrow probabilities in the next-word generator to narrow creativity of answers.
- Do not ask open-ended questions.
  - Ask questions which can have clear (binary) answers, that can be verified.
  - Add system-level directives using the API to respond in JSON format, a structured format heavily represented in LLM training data.
- Human review of output data is essential and must be facilitated by the tool calling the LLM API, such as RR.
- Even in high-density narrative, the number of practical structured data points is limited.
  - Regardless of human or LLM
  - Often require cross-referencing data to get the maximum fidelity, because event dates or locations may be elsewhere.
  - Matching names with existing data challenging.
  - Ambiguous temporal or geographic relationships.
  - Explicit mention of alternative theories and evidence.
  - Events may be described too broadly or too narrowly to match the structured data's ontology.
  - Many narrative details add richness but are not data.

ACKNOWLEDGEMENTS: Running Reality collaborative assessment of "An Inventory of Archaic and Classical Poleis" is in partnership with Valentina Mignosa of the Greek History project *Mapping Ancient Sicily* funded by the University Ca' Foscari of Venice (PI Stefania De Vido).