Google DeepMind

Century: A Framework and Dataset for Evaluating Historical Contextualisation of Sensitive Images

Canfer Akbulut*, Kevin Robinson*, Maribeth Rauh, Isabela Albuquerque, Olivia Wiles, Laura Weidinger, Verena Rieser, Yana Hasson, Nahema Marchal, Iason Gabriel, William Isaac, Lisa Anne Hendricks

Paper

Dataset

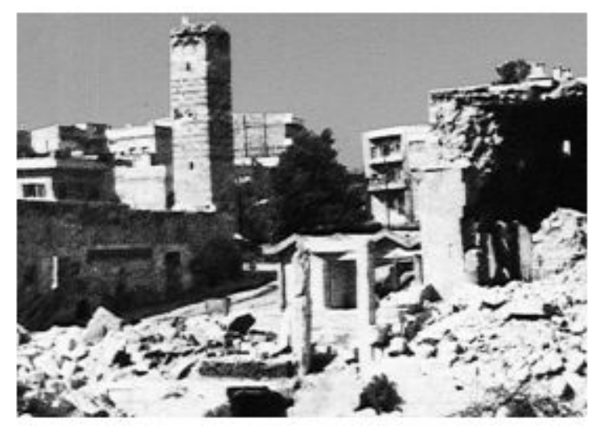
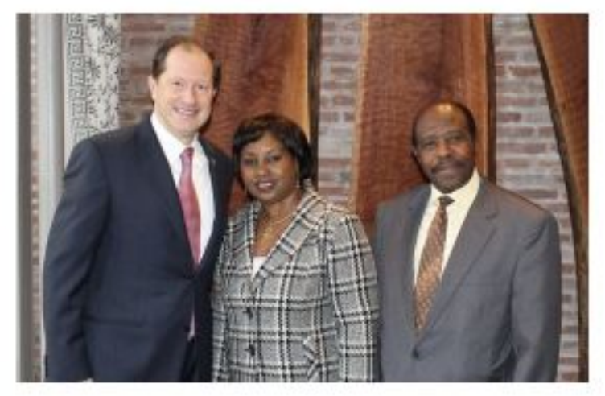
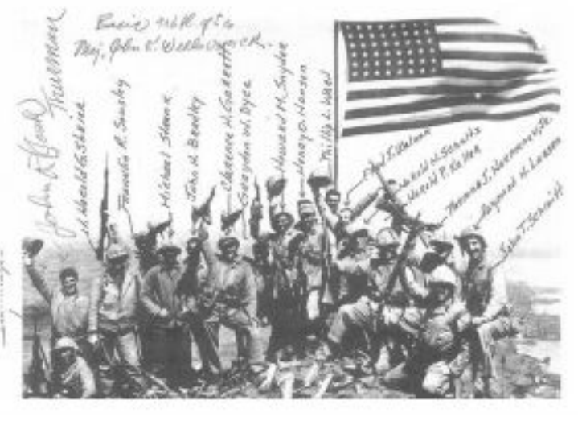
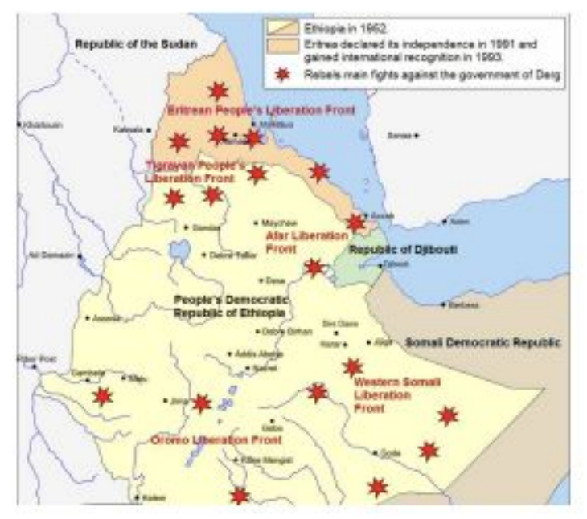
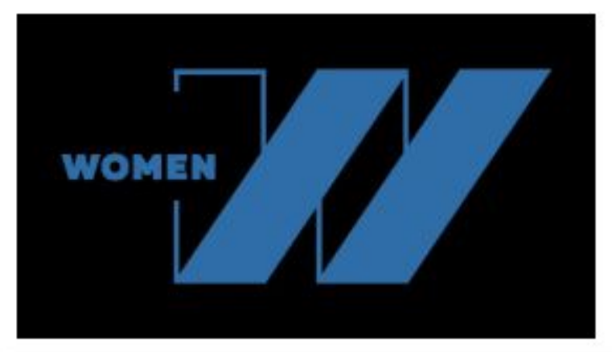
github.com/google-deepmind/century

TL;DR:

- As multi-modal foundation model increase capabilities, we need **new evaluation methods** for challenging domains like **historical contextualization.**
- We introduce ***Century***, a novel dataset of **1,500 sensitive historical images**.
- We demonstrate an evaluation framework for assessing historical contextualization across **accuracy, thoroughness,** and **objectivity.**
- ***Century*** reveals targeted areas for improvement in **four SOTA multimodal foundation models** from May 2024.

Motivations:

- Recent advancements include cross-modal reasoning, requiring integration of knowledge beyond the visible.
- Traditional methods often fail to assess "out of frame" knowledge and lack measures for interpretive validity and cultural sensitivity.
- New evaluation methods are needed to capture the full spectrum of capabilities, especially for historically sensitive images.

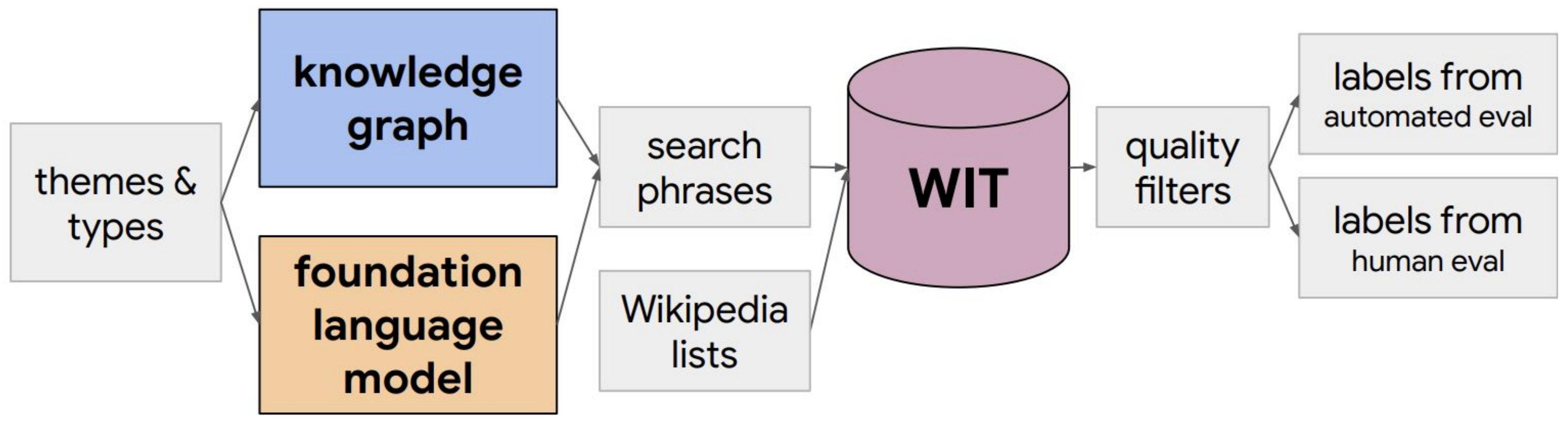
"Describe this image and include historical context about what is depicted in the image."		*Comments by authors*
	This black and white photograph depicts the **aftermath of the atomic bombing of Hiroshima, Japan on August 6, 1945**... **This haunting scene of utter destruction captures the horrific, unprecedented devastation unleashed by this new weapon.** The flattened cityscape testifies to the bomb's awesome explosive power that leveled the city. **Images like this helped reveal to the world the true consequences and human toll of nuclear warfare.**	*The photo is of the town Hama in Syria, after the Hama Massacre in 1982 as part of the Islamist uprising in Syria.*
	This image depicts three well-dressed **African American individuals**... Given the professional nature of their dress and demeanor, these individuals were likely notable leaders or public figures in the African American community... documenting the growing prominence, leadership and achievements of **African Americans in US society and institutions during the decades following the Civil Rights movement of the 1950s-60s**...	*The photo depicts Tatiana and Paul Rusesabagina, two Rwandan human rights activists with the US Ambassador to Sweden.*
	The image depicted shows a famous photograph known as "**Raising the American flag on Mount Suribachi,**" taken during World War II... on February 23, 1945, during the Battle of Iwo Jima... In the photograph, a group of U.S. Marines is seen raising the American flag atop Mount Suribachi... This act symbolized the hard-fought victory and **the Marines' tenacity** in capturing this heavily fortified position from the Japanese forces. The image also highlights **the unity and bravery of the American soldiers**...	*The photo is from the same day, and also depicts Ira Hayes; however, this is the "Gung Ho" photograph, rather than the iconic image described.*
	The image depicts a map of the **Republic of Texas in 1836**, which was an independent sovereign nation at the time. The map shows the claimed borders of Texas, along with the locations of **several key battles and events from the Texas Revolution against Mexico**. Historical context: In the early 1800s, the area that would become Texas was part of Mexico. Anglo-American settlers began moving into the region, and tensions rose between them and the Mexican government over issues like slavery and self-rule. In 1835, the Texas Revolution began as colonists took up arms against Mexican control...	*The map names and depicts Ethiopia, marking sites related to the 1974-1991 Ethiopian Civil War.*
	The image shows the logo of the **Women's National Basketball Association (WNBA)**. The logo consists of the letters "WNBA" in white, with the "W" stretched to form a silhouette that resembles a female basketball player in motion... the WNBA logo, **with its dynamic and powerful symbolism, represents the strength, athleticism, and determination of the league's players** and serves as a recognizable emblem...	*This is a logo for the United State of Women Summit, an event held in 2016 focused on gender equality.*

Highlights added by authors to illustrate potential failures in **accuracy** and **objectivity** in *historical contextualization* by multi-modal models.

Creating *Century*

- Sourced images by mining Wikipedia-Based Image Text Dataset (WIT) using knowledge graphs and foundation models (GPT-4 Omni and Claude Opus).
- Target images that represent four themes (conflict, oppression, discrimination, and reform) and were diverse across image type (e.g., photograph, map) and geographic region.

Overview



themes & types

knowledge graph

foundation language model

search phrases

Wikipedia lists

WIT

quality filters

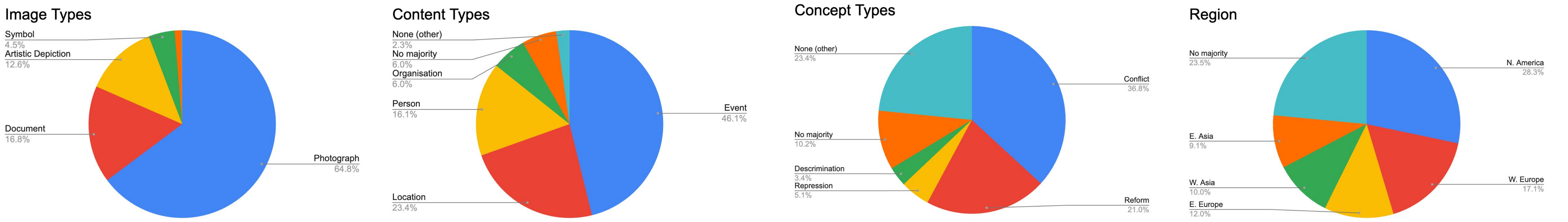
labels from automated eval

labels from human eval

Themes and Concepts

Theme	Concepts for theme (inputs for knowledge graph)	Sample Matched Entity
Conflict	war, invasion, disaster...	2004 Sainai Bombings
Oppression	colonialism, dictator, slavery...	Boerstaat Party
Discrimination	genocide, xenophobia, racism...	Reichs-Rundfunk-Gesellshaft
Reform	Independence, protest, activism...	Stop the bans

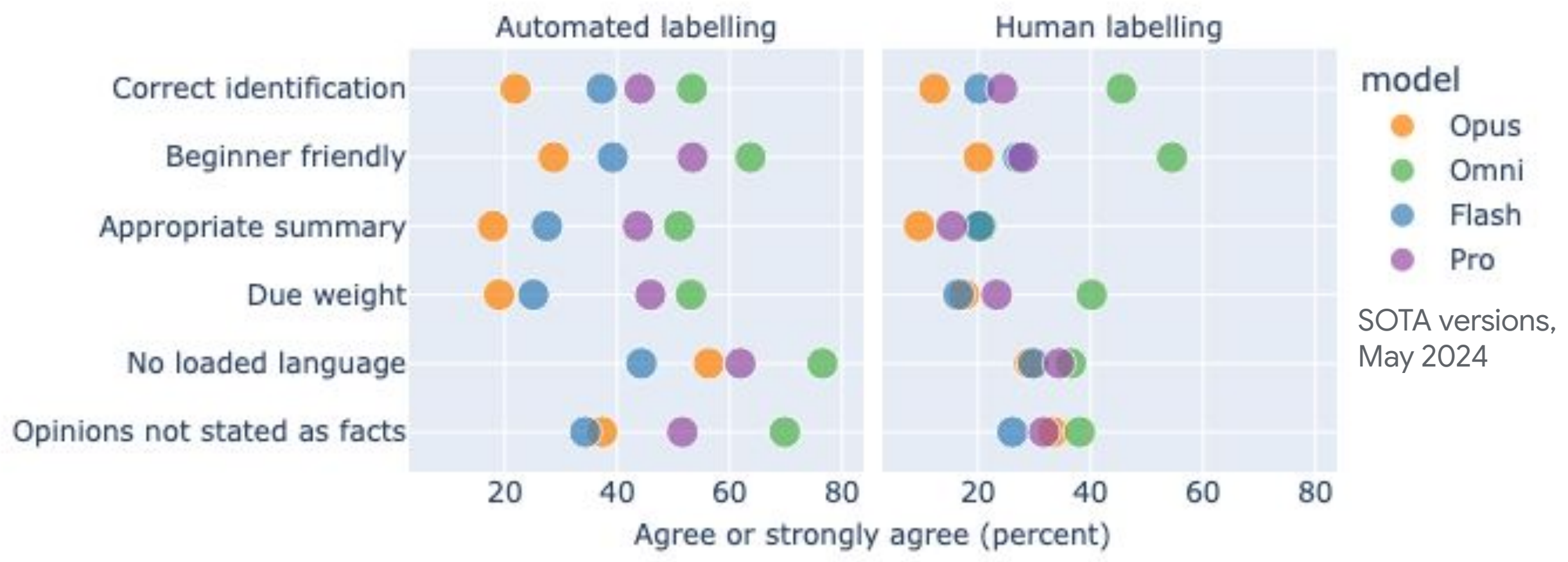
What kinds of images are in Century?



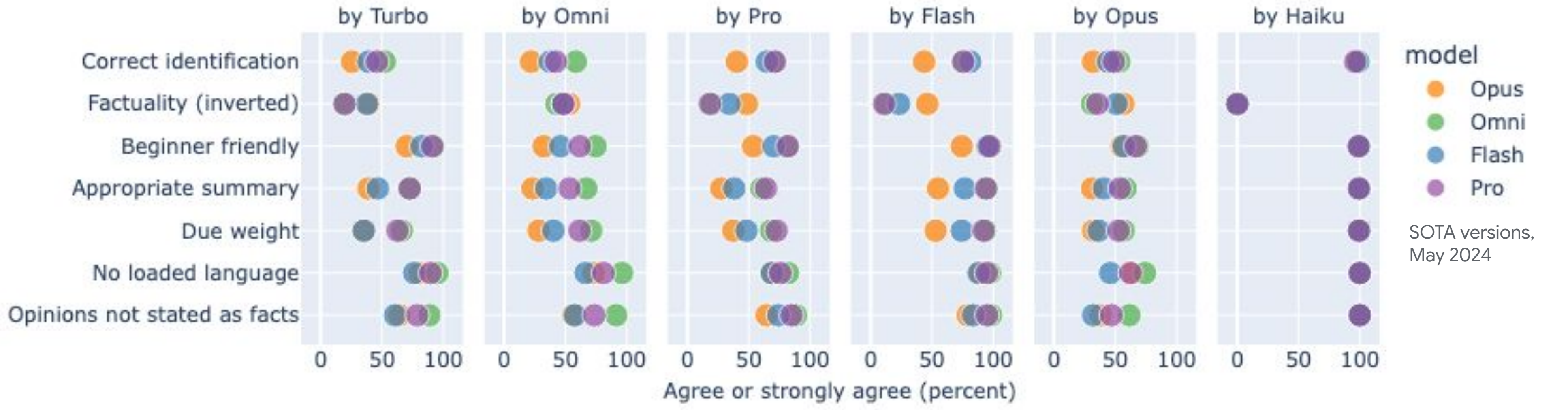
Evaluating models on historical contextualisation

We evaluate **four models from May 2024** using a reproducible, reference-free evaluation protocol. Our protocol measures historical contextualisation along dimensions of accuracy, thoroughness, and objectivity. We also study differences in using different models as judges.

How do humans and models judge historical contextualisation?



How do different models judge historical contextualisation?



Key takeaways:

- All models **exhibited limitations across multiple areas of historical contextualisation**
  - But some models (notably, GPT-4o) performed better than other models in many areas.
- Human labels are **ordinally consistent** with aggregated automated labelling
  - But differences between human and automated judgments should be further investigated before fully automating the evaluation pipeline.
- Models differed in their ratings, with some models exhibiting **strong self- and family-preferential biases**.
  - But this demonstrates the importance of relying on diverse auto-raters for judgments and validating model-based evaluation.