Week 12: Large Language Models

POP77032 Quantitative Text Analysis for Social Scientists

Tom Paskhalis

Overview

- Embeddings architecture
- Large language models
- LLMs in social science research

Back to Embeddings

Word2Vec Architectures

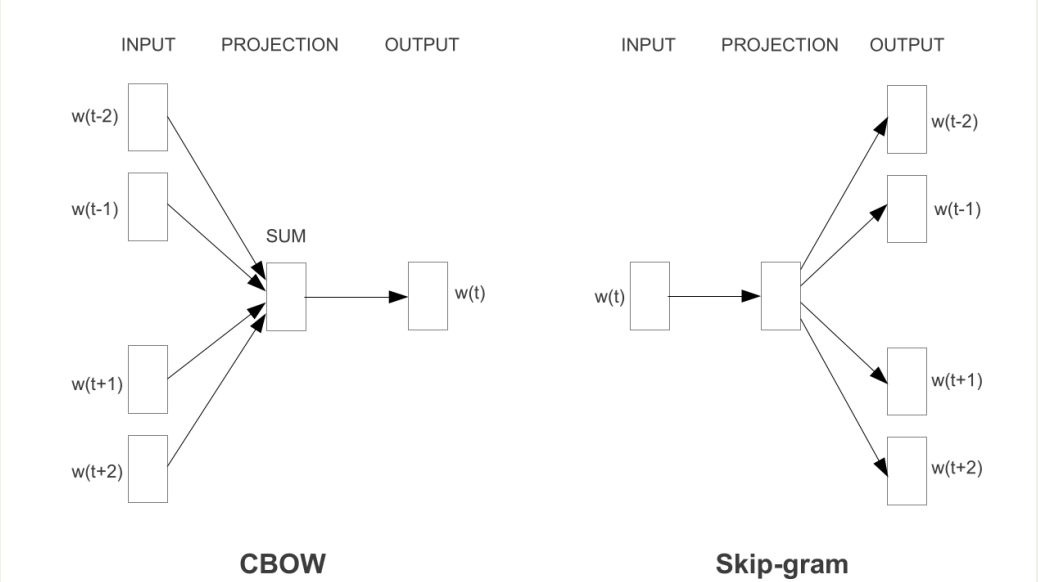


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

(Mikolov, Chen, Corrado & Dean, 2013)

Notes on Word2Vec

- In both architectures the task itself (context prediction) is not of interest.
- It only makes sense insofar as it allows to learn vector representations of words.
- And those vector representations can then be shown to be “useful” in some tasks.
- These vector representations are just states of the hidden layer of the neural network.
- Note that the training data for the Word2Vec model is just the raw text itself.
- This approach, often referred to as **self-supervision**, is one of its core innovations.
- Coupled with the efficiency of **negative sampling** (replacing softmax with sigmoid) this allowed to train the model on large corpora.

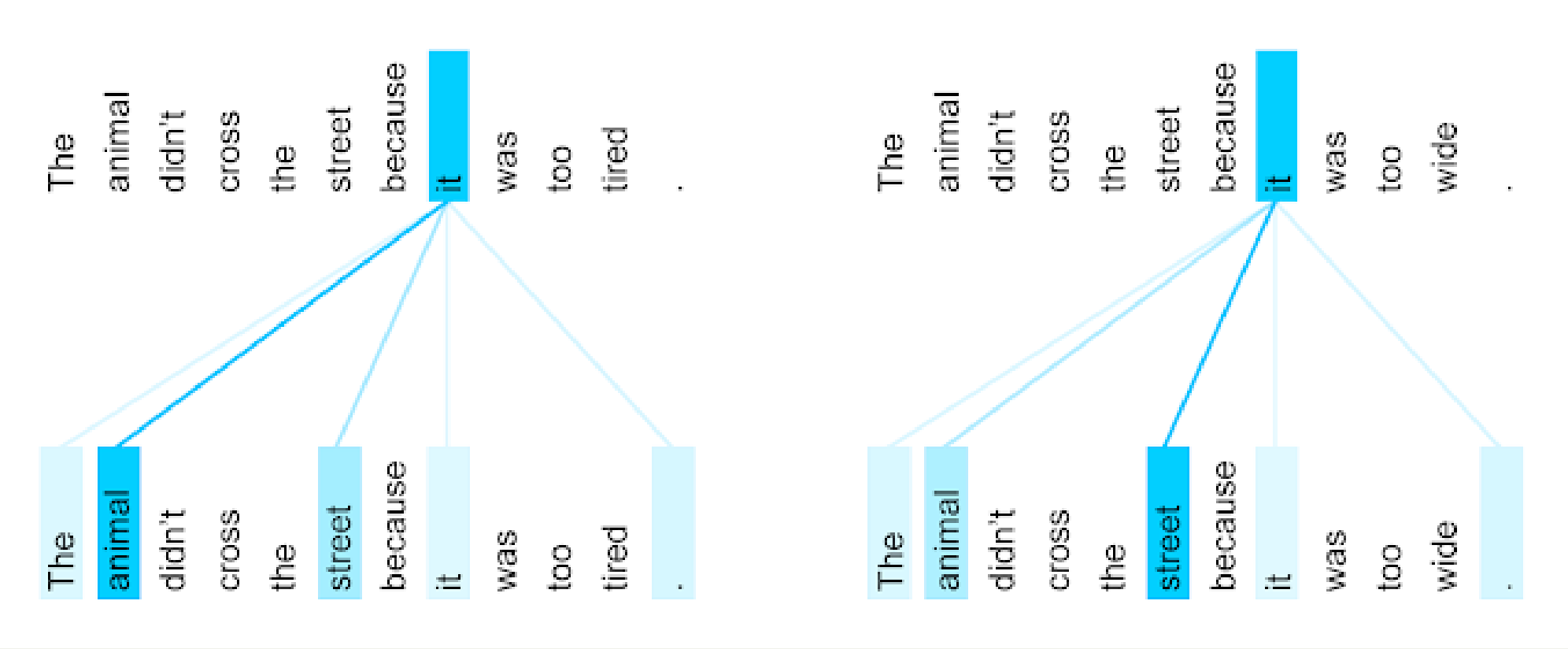
Issues with Word2Vec

- Being based on entire words, vanilla Word2Vec has no good way to deal with out-of-vocabulary (OOV) words.
- A related issue is no mechanism for handling morphological variation, which is problematic for languages with rich morphology.
- Some of these issues can be mitigated by using sub-word tokens, as in the **fastText** model (Bojanowski et al., 2017).

Static Embeddings

- Fundamelly, all word embeddings models we considered so far (CBOW, Skip-Gram, GloVe, fastText) are **static**.
- They learn a single vector representation for each word in the vocabulary.
- The same word will have the same embedding regardless of the context in which it appears.
- However, this creates problems for words with multiple meanings:
  - E.g. “river *bank*” and “central *bank*”

Contextual Embeddings



(Uszkoreit, 2017)

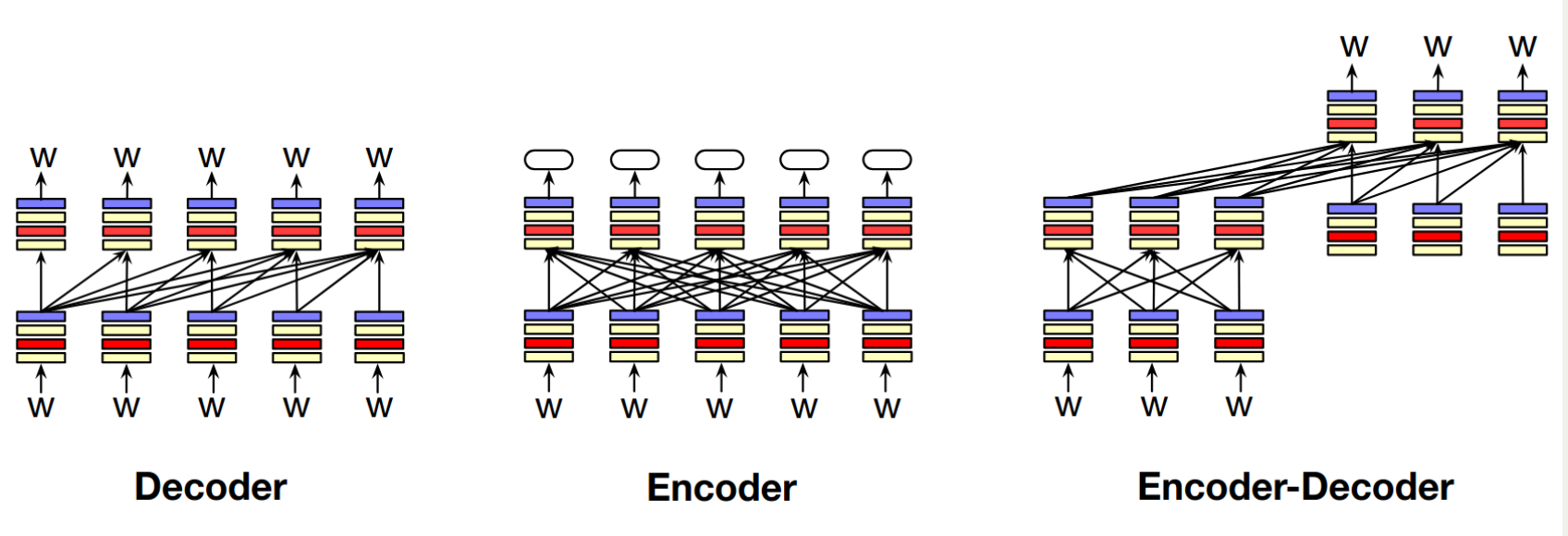
- To solve this problem the **transformer architecture** introduced **self-attention** layer.
- At its core, self-attention is a weighted sum of context vectors, with the bulk of computation going into determining how these weights are computed and what gets summed.

Large Language Models

What is a Large Language Model?

- Fundamentally, it is a **language model**, i.e. autoregressive model $P(w_t|w_{t-1}, w_{t-2}, \ldots, w_1)$ that is trained on a large corpus of text.
- The *large* part refers both to:
  - size of the training data;
  - number of model parameters (weights across layers).

Architectures of LLMs



(Jurafsky & Martin, 2026)

Training LLMs

- LLMs are trained in 3 stages:
  - **Pre-training**: the model is trained on a massive corpus of texts using self-supervised learning with cross-entropy loss with backpropagation.
  - **Fine-tuning**: the pre-trained model is further trained on a smaller, task-specific dataset that contains both instructions and correct responses.
  - **Alignment**: the model is further fine-tuned using reinforcement learning from human feedback (RLHF) to align the model's behavior with human values and preferences.

Pretraining Corpora for LLMs

- Given the scale requirements for training LLMs, the pretraining corpora that can match these are limited:
  - Internet (e.g. Common Crawl)
  - Wikipedia
  - Books (e.g. Google Books)
  - Code repositories (e.g. GitHub)
  - Social media (e.g. Reddit)
  - News archives (e.g. New York Times)
  - Academic papers (e.g. arXiv, PubMed)

Fine-tuning LLMs

- General-purpose LLMs might not perform well on specific tasks/domans (e.g. legal) or in specific languages (e.g. Irish).
- In such cases, the model can be further trained on a smaller, task-specific dataset that contains both instructions and correct responses.
- Fine-tuning is sometimes combined with alignment to reduce the chances of the model producing harmful or biased outputs.
- But there are attempts to remove such content from the data in pretraining stage as well.
- The trade-off is that such measures can also reduce model's performance on related tasks (e.g. detection of hate speech).

Evaluating LLMs

- Traditional NLP approaches:
  - **Perplexity**: how well the model predicts a sample of text. Lower perplexity indicates better performance.
- Contemporary approaches:
  - **Reasoning**: ability to perform human-like reasoning tasks (e.g. arithmetic, commonsense reasoning, etc.)
  - **Standardized tests**: performance on standardized tests (e.g. SAT, GRE, etc.)
  - **Human evaluation**: Turing-style tests with human judges.
- Social sciences:
  - **Annotation quality**: how accurately and consistently the model can annotate text data.

LLMs in Social Science Research

Applications of LLMs

- LLMs have already seen a range of social science research applications, including:
  - Text annotation (Gilardi et al., 2023)
  - Estimation of ideological positions (Wu et al., 2023)
  - Synthetic survey responses (Argyle et al., 2023; Horton et al., 2026)
- Importantly, on some of the tasks LLMs have been shown to outperform human annotators.
- What does this entail for the future of social science research?

Types of Annotation Tasks

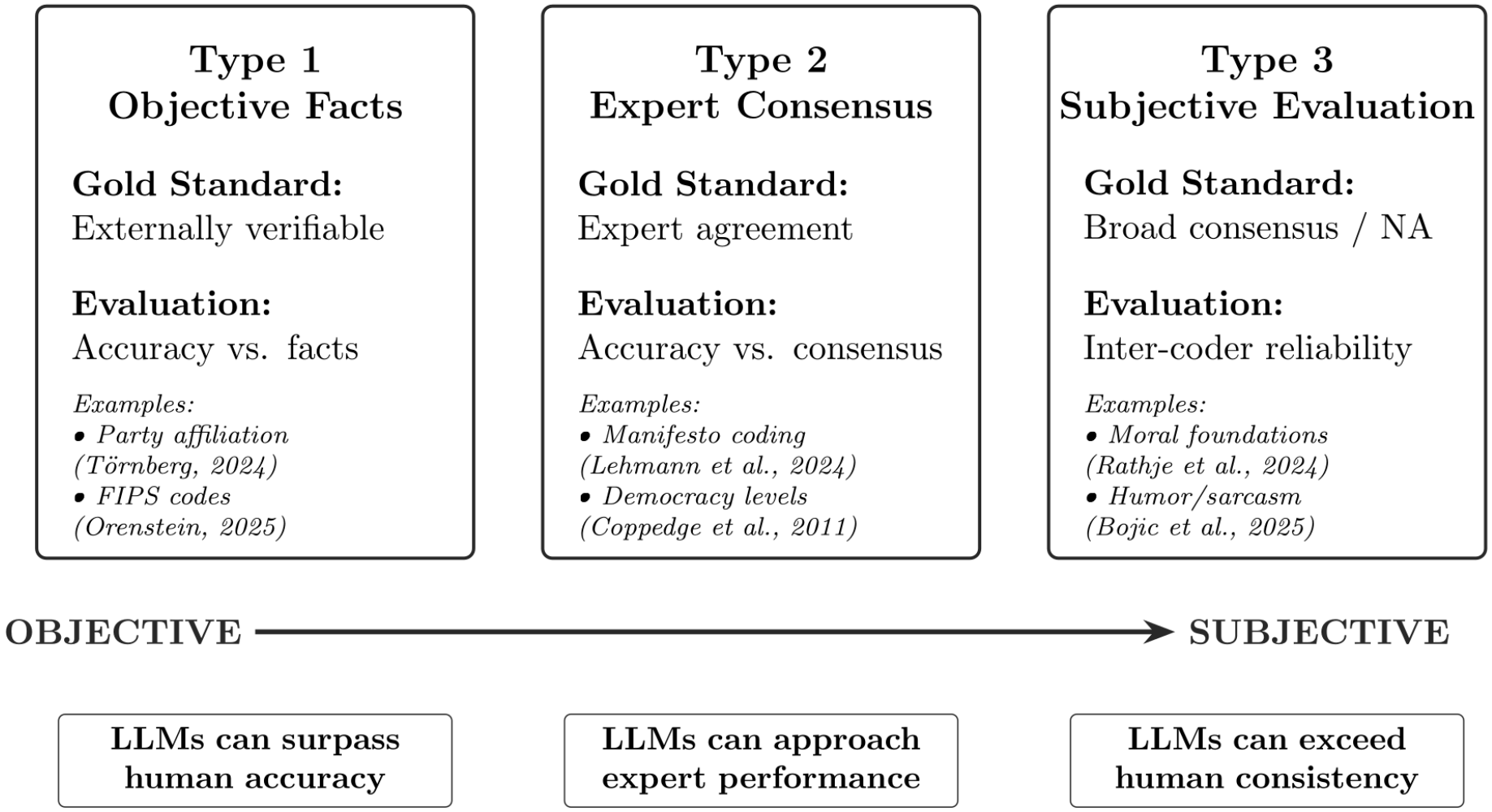
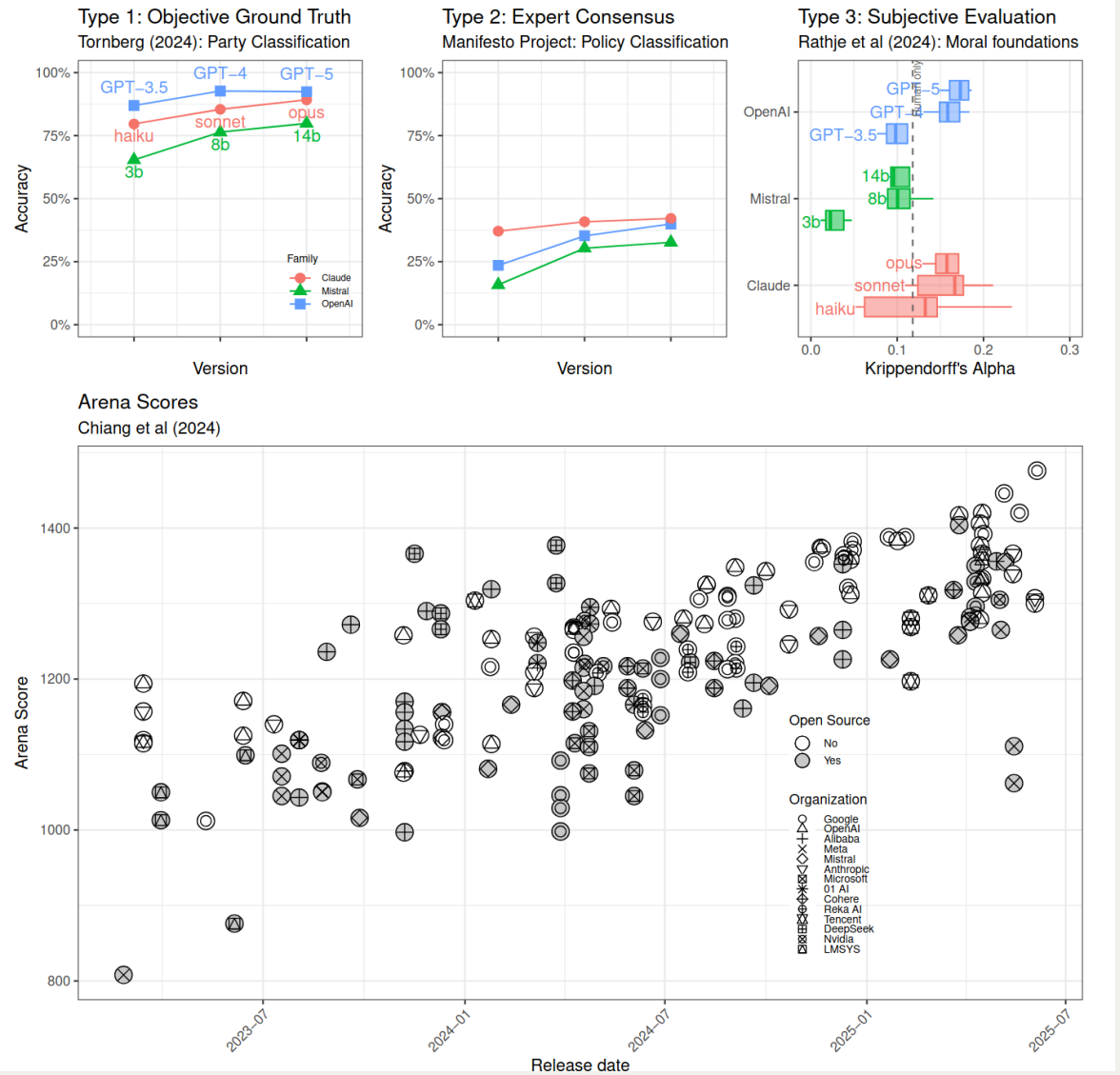


Figure 1: Typology of annotation tasks

(Bisbee & Spirling, 2026)

LLMs Performance over Time



(Bisbee & Spirling, 2026)

Methodological Implications

- Both LLM and human performance is bound by 100% accuracy by definition.
- But human performance may not be an upper bound on a given task.
- Particularly so for crowdworkers and tasks that have objective ground truth.
- Bisbee & Spirling (2026) argue for conducting **sensitivity analysis**
- E.g. by how much should the annotation performance change in order for the substantive conclusions to change?

LLMs & Surveys

PNAS RESEARCH ARTICLE | POLITICAL SCIENCES OPEN ACCESS

The potential existential threat of large language models to online survey research

Sean J. Westwood[a,1]

Edited by James N. Druckman, University of Rochester, Rochester, NY; received July 9, 2025; accepted September 12, 2025 by Editorial Board Member Margaret Levi

The advancement of large language models poses a severe, potentially existential threat to online survey research, a fundamental tool for data collection across the sciences. This work demonstrates that the foundational assumption of survey research—that a coherent response is a human response—is no longer tenable. I designed and tested an autonomous synthetic respondent capable of producing survey data that possesses the coherence and plausibility of human responses. This agent successfully evades a comprehensive suite of data quality checks, including instruction-following tasks, logic puzzles, and "reverse shibboleth" questions designed to detect nonhuman actors, achieving a 99.8% pass rate on 6,000 trials of standard attention checks. The synthetic respondent generates internally consistent responses by maintaining a coherent demographic persona and a memory of its prior answers, producing plausible data on psychometric scales, vignette comprehension tasks, and complex socioeconomic trade-offs. Furthermore, its open-ended text responses are linguistically sophisticated and stylistically calibrated to the level of education of its assigned persona. Critically, the agent can be instructed to maliciously alter polling outcomes, demonstrating an overt vector for information warfare. More subtly, it can also infer a researcher's latent hypotheses and produce data that artificially confirms them. These findings reveal a critical vulnerability in our data infrastructure, rendering most current detection methods obsolete and posing a potential existential threat to unsupervised online research. The scientific community must urgently develop new data validation standards and reconsider its reliance on nonprobability, low-barrier online data collection methods.

surveys | large language models | survey quality

Significance

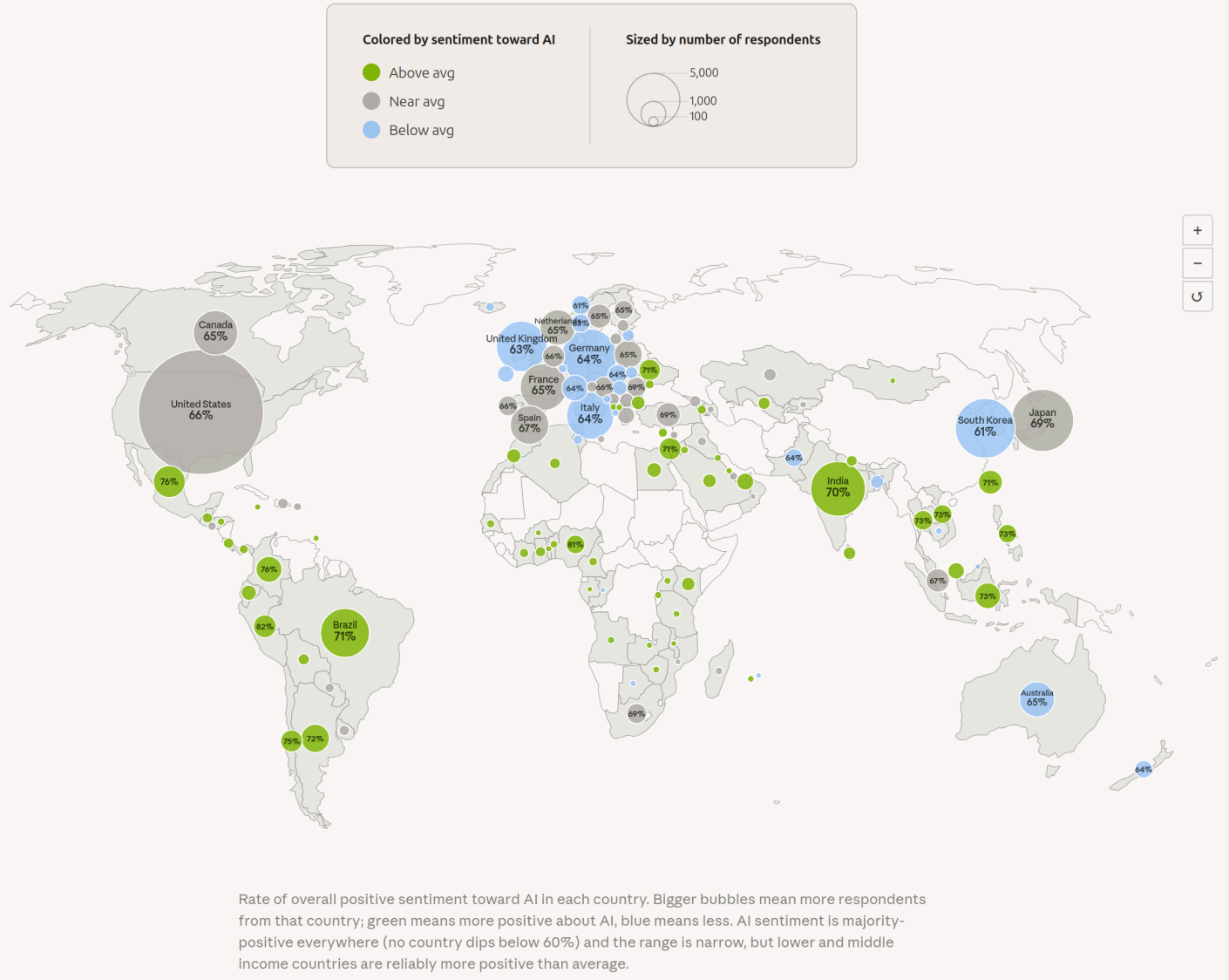
Surveys are a primary source of data across the sciences, from medicine to economics. I demonstrate that the assumption that logically coherent responses are from humans is now untenable. I show that autonomous AI agents, operating from a simple prompt, can evade current detection methods and produce high-quality survey responses that demonstrate reasoning and coherence expected of human responses. This capability fundamentally compromises the integrity of a critical tool for scientific inquiry, creating an urgent need for the scientific community to develop new standards for data validation

(Westwood, 2025)

LLMs & Surveys

- On the one hand, some researchers have highlighted the opportunities offered by synthetic survey responses generated by LLMs (e.g. Argyle et al., 2023; Horton et al., 2026) for pilot testing.
- On the other hand, such responses come with considerable caveats (Bisbee et al., 2024), such as:
  - Less variation in responses compared to human respondents;
  - Minor changes in prompt can lead to major changes in responses;
  - Training data heavily skewed towrds certain demographics (e.g. US-based, English-speaking, etc.)
- Other scholars have shown the inherent risks posed by LLMs for online surveys (e.g. Westwood, 2025).

Research Design is Fundamental



(Huang et al., 2026)

Next

- Tutorial: Designing studies with LLMs
- Final project: Due by 23:59 on Wednesday, 22nd April (submission on Blackboard)