

When More Data Is Not Enough: Language as Ideological Memory in Gendered Language Models

Abstract

In language modeling, increased training data is often assumed to yield better alignment with human social attitudes; however, this assumption overlooks the historical and ideological structure embedded in language itself. Recent work suggests that continued pretraining on domain- or time- specific corpora can improve language models' alignment with human social attitudes. Yet, it remains unclear whether such improvements are primarily driven by increased data quantity or by the historical-ideological structure of the data itself. The study aims to disentangle these effects by continue-pretraining BERT on decade-specific English corpora between 1910s and 2010s and evaluating model alignment with human gender survey and real-world occupational gender distributions.

In contrast to the scaling hypothesis, our results demonstrate that the improvements in BERT's alignment with human gender attitudes through continue-pretraining cannot be explained by data scale alone. Interestingly, we find no linear association between gains of alignment and the size of training-data: models trained on cumulatively expanding corpora (from the 1950s to the 2010s) do not exhibit progressively stronger correlations with the 1950s survey. Rather, their performance closely resembles that of models trained on individual decades, differing primarily by a systematic temporal lag relative to major historical events. Moreover, maximal alignment does not necessarily occur when the training corpus temporally matches the survey year. All these patterns show that the gender ideology encoded in language does not evolve linearly over time, but follows a non-monotonic historical process marked by reversals, oscillations, and delayed linguistic expressions in response to significant societal and political events.

To further examine what language models capture, we analyze correlations between model representations and occupational gender composition across decades. We find that alignment curves derived from occupation data in the 1950s, 1970s, and 1990s exhibit nearly identical shapes across all pretrained models, differing primarily by an overall shift in magnitude. This pattern suggests that occupational gender structure constitutes a long-run, structurally stable dimension that is persistently reflected in language, regardless of short-term variation in labor-force participation. The consistency of these curves is compatible with the view that language encodes gender ideology in a temporal and smoothed manner, emphasizing durable social organization over contemporaneous attitudinal change.

Taken together, these findings suggest that continued pretraining primarily strengthens language model's alignment with long-lasting ideological frameworks embedded in texts, rather than simply reflecting increased data or short-term attitudinal change. The paper highlights the necessity of temporally framed and historically oriented approaches to the assessment, interpretation, and alleviation of social bias in language models.

Key Words: Historical language modeling; Gender ideology; Continued pretraining; Social bias in language models