Abstract

This project investigates computational prescriptivism as an algorithmic enforcement of linguistic standardization in NLP-based summarization systems, with particular focus on legal deposition transcripts. Through empirical analysis of real legal deposition excerpts and curated synthetic examples across six summarization models (BART, Pegasus, T5/Flan-T5, Lead-2, TextRank, and GPT-3.5), this study demonstrates that large language models routinely erase pragmatic features that are disproportionately associated with speakers of non-standard English dialects and women.

The research reveals a consistent hierarchy of feature vulnerability, with disfluency markers displaying the highest erasure rates, followed by hedges and modal expressions; temporal and conditional markers are more frequently preserved. Critically, the analysis documents certainty inflation, where hedged statements are transformed into categorical claims. These distortions emerge from specific architectural mechanisms: Pegasus's gap-sentence generation objective leads to hallucination; BART's denoising autoencoder treats natural disfluencies as corruption; and T5's span corruption task causes attention mechanisms to skip low-confidence tokens.

To address these issues, this work introduces the Pragmatic Distortion & Certainty Index (PDCI). This conceptual evaluation framework quantifies stance distortion through two components: the Certainty Inflation Index and Pragmatic Distortion Score. The framework reveals considerable variation across models. While alignment-tuned models are more equipped to conserve linguistic nuance through prompting, explicit preservation instructions cannot fully override learned biases.

The findings reveal that computational prescriptivism operates through training objectives optimized for edited, standardized text rather than spontaneous discourse, creating systematic disadvantages for speakers whose language patterns deviate from dominant norms. In legal contexts, where credibility assessment is informed by the pragmatic features that models are inclined to reformat or eliminate, this algorithmic erasure poses serious risks to procedural fairness. The study suggests that current NLP systems function as algorithmic language authorities that enforce standardization through technical rather than social mechanisms, with implications extending beyond law to any domain where AI mediates human communication and credibility matters.