Can LLMs Retrieve Variation? A Corpus-RAG Study of Native-Speaker Norm Drift in English as a Lingua Franca Dialogue

Seongyong Lee, University College London, seongyong.lee@ucl.ac.uk
Jaeho Jeon, Korea National University of Education. jeon@knue.ac.kr
Yonhee Chung, University College London, yonhee.chung.25@ucl.ac.uk

Research reports that large language models (LLMs) often produce standardized English even when asked to support English as a Lingua Franca (ELF) (Author, 2026). Recent sociolinguistic work argues that LLMs model particular language varieties rather than language in general (Grieve et al., 2025), whereas AI writing systems may preserve meaning while erasing markers of World Englishes (Navneet et al., 2026). Building on Author (2026) in which ELFA-based retrieval-augmented generation (RAG) improved ELF-like responses but gradually drifted back toward native-speaker norms, this study explored whether corpus-RAG could make sociolinguistic variation retrievable and sustainable in LLM-generated ELF dialogue.

It compares five conditions, adopting GPT 5: (1) baseline LLM, (2) ELF instruction-only prompting, (3) vanilla RAG using sentence/turn chunks from ELFA, (4) metadata-aware RAG using event type and speaker-role information, and (5) strategy-coded RAG that uses annotated repair, clarification, repetition, accommodation, and negotiation episodes. 60 ELFA interactional episodes were sampled and annotated for pragmatic moves. Each condition completed four tasks: responding to an ELF utterance, repairing a trouble source, continuing a three-turn dialogue, and reformulating an utterance without unnecessary correction.

Regarding outputs analysis, corpus-functional benchmarks compared generated responses with attested pragmatic moves in similar ELFA episodes. Expert coding assessed strategy appropriateness, non-deficit orientation, and interactional authenticity. Quantitative indices included ELF Strategy Presence, Standardization Pressure, Errorization Rate, Feature/Function Retention, Corpus-groundedness, and Norm Drift Slope.

Results show that (1) instruction-only prompting improved ELF likeness but maintained native-speaker norms; (2) vanilla RAG increased corpus-grounded cues but fragmented interactional sequence; (3) strategy-coded/metadata-aware RAG better maintained repair, clarification, and accommodation while reducing overcorrection. We argue that sociolinguistic representativeness in LLMs is not achieved only by adding diverse data but also depends on how corpus evidence is segmented, annotated, retrieved, and recontextualized.

References

Grieve, J., Bartl, S., Fuoli, M., Grafmiller, J., Huang, W., Jawerbaum, A., Murakami, A., Perlman, M., Roemling, D., & Winter. B. (2025). The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence, 7*, 1472411. https://doi.org/10.3389/frai.2024.1472411

Lee, S., Jeon, J., McKinley, J., & Rose, H. (2025). Generative AI and ELT: A Global Englishes perspective. *Annual Review of Applied Linguistics, 59*(1), 49-74. https://doi.org/10.1017/S0267190525100184

Navneet, S. K., Chandra, J., & Zhang, Y. (2026). When AI writes, whose voice remains? Quantifying cultural marker erasure across World English varieties in large language models, *arXiv, 2602*, 22145. https://doi.org/10.1145/3772363.3799085