

Nordic AI Meet 2026

P3: LLMs, Language, Evaluation, Privacy & Trustworthy AI

2. Stílmæliprófið: A Stylometric Benchmark for LLM-Generated Icelandic

Authors: Esther Ýr Þorvaldsdóttir¹

Affiliations: 1. Faculty of Icelandic and Comparative Cultural Studies, University of Iceland, Iceland

Abstract: For low-resource languages like Icelandic, the quality of text generated by large language models varies widely across models. Fluency is not guaranteed and even if a model is grammatical it may not generate well across registers. Measuring style, not just correctness, is as much a linguistic question as an engineering one. The quantitative nature of this study is the prism that breaks apart the prose and reveals the uncanniness that Icelanders can feel while reading generated text, but can't quite put their finger on. I present the first benchmark that studies the linguistic style of LLM-generated Icelandic prose, adapting the paired-continuation methodology of Milička et al. (2025) to a morphologically rich, low-resource setting.

Six models were prompted with the first ~1000 words of an authentic human sample, most from the Icelandic Gigaword Corpus (IGC), written in four different registers, and the models' job was to copy the register when continuing the text. The continuation was compared to the human original across 11 linguistically motivated features, including syntactic depth, vocabulary coverage against the BÍN morphological database, LIX readability, and lexical diversity (MTLD). Syntactic features are computed from constituency parses (IceConParse), making the benchmark sensitive to grammar-level style rather than surface lexical statistics alone. Texts span three register types from the IGC (academic, news, and blog), plus previously unpublished fiction. Deviations were normalized against the original human sample and aggregated into a single style score.

I evaluated the models via their chat interfaces: Claude Sonnet 4.6, GPT-5, Gemini 3 Thinking, DeepSeek V3.2, and two Mistral Le Chat variants. Gemini 3 Thinking imitated the human style most closely, while the Le Chat models underperformed. Lexical diversity was the strongest discriminator between human and generated text across all models. Most models generated less lexically varied Icelandic than humans but more syntactically embedded; sparse vocabulary but bushier syntax (Gemini being the exception by overshooting on vocabulary richness). Even competent models deviate, but no single feature identifies the deviation. As feature measures were added, the uncanny quality became easier to locate. Results are published on a live public leaderboard (Hugging Face Spaces) with user-adjustable feature weights, allowing researchers to prioritize features they consider important in Icelandic style. Gemini ranks first across every weighting I tried, and the Le Chat models last. Rank order in the middle shifts based on weights; that's where things get linguistically interesting.

My plan is to extend the benchmark with mid-sentence continuation prompts, fully contamination-free human language samples, API-level prefill, and thematic multi-feature dimensions.

This work is supervised by Steinunn Rut Friðriksdóttir.