Do Large Language Models Style-Shift? Register-Conditioned Stylistic Fronting in AI-Generated Icelandic

Anton Karl Ingason
University of Iceland
antoni@hi.is

Johanna Mechler
University of Iceland
mechler@hi.is

Lilja Björk Stefánsdóttir
University of Iceland
lbs@hi.is

Abstract

We examine whether large language models (LLM) encode sociolinguistic knowledge of formal and informal register in a low-resource language, Icelandic, finding that register sensitivity is present in the two systems tested (GPT-5.5 and Gemini 3.1 Pro). Our findings are based on a 2×2 persona design with elicited production (n=1,200), manually coded for Stylistic Fronting, a property of formal language. The direction and size of their style shift are not detectably different, but they differ sharply in overall rate; against a matched human comparator, only GPT-5.5 produces a persona-appropriate rate. Neither system reproduces the gender pattern documented for human speakers. Thus, register sensitivity does not guarantee that output rates are appropriate to the persona the model was instructed to adopt, an issue that may reflect limited training data for low-resource languages.

1 Introduction

In recent research, Large Language Models (LLMs) are increasingly used as proxies for human speakers (e.g., Argyle et al., 2023), and their sociolinguistic properties are under ongoing investigation (e.g., Hofmann et al., 2024). A key open question in this respect is whether and how such systems encode sociolinguistic knowledge of languages that are underrepresented in their training data. Icelandic is an interesting test case for exploring this issue, since it is low-resource in terms of Language Technology and has a well-developed sociolinguistic literature on syntactic variation. This combination allows us to examine LLM-generated Icelandic in the context of what we know about human language variation.

We examine Stylistic Fronting (SF) (Ingason and Mechler, 2025; Mechler et al., 2025), a word order phenomenon that is well documented and sensitive to register. By controlling the syntactic context, we get a binary contrast in SF that is easy for an expert human to annotate, thus being ideal for controlled elicitation.

While LLM performance has been studied on certain tasks in Icelandic (Ármannsson et al., 2025b,a; Ingimundarson et al., 2026), no previous study has examined whether and how LLMs encode sociolinguistic knowledge about Icelandic that allows them to distinguish between formal and informal registers. We test two independently developed systems (GPT-5.5 and Gemini 3.1 Pro) and show that they produce significantly more SF in the formal conditions than the informal ones, thus register sensitivity is present, and not vendor-specific. The direction and size of the shift are not detectably different between the systems, but their overall usage rate differs sharply. Against a matched human comparator averaging 90.4% SF, only GPT-5.5 falls within that speaker's attested range (81.8%), while Gemini 3.1 Pro produces 34.0%, so register sensitivity does not guarantee a persona-appropriate rate. We find that neither system reproduces the gender pattern documented for human speakers (see Section 2.1).

The paper is structured as follows. Section 2 outlines the background, then Section 3 discusses our methods. We present our results in Section 4. Section 5 provides discussion and Section 6 concludes.

2 Background

2.1 Sociolinguistic variable: Stylistic Fronting

A sociolinguistic variable is defined as two or more ways of expressing the same meaning, where choice between variants is conditioned by social and linguistic factors, not by truth conditions (Labov, 1972). Here, we focus on SF as a syntactic sociolinguistic variable. SF is a word-order phenomenon in Icelandic clauses with subject gaps.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 140–149
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_016

This includes some relative clauses where a non-finite main verb may optionally precede a finite auxiliary (Maling, 1980; Holmberg, 2006; Thráinsson, 2007; Angantýsson, 2011). The following examples show the two word orders in such a clause. Example (1) shows the auxiliary-first order (no SF); Example (2) shows the non-finite-first order (SF). Using either variant does not affect the truth condition of the sentence.

(1) *Varðandi* *það* *[$_{CP}$* *sem* ***var*** ***sagt*** *hér]*
regarding it that **was** **said** here
'Regarding what was said here' [no SF]

(2) *Varðandi* *það* *[$_{CP}$* *sem* ***sagt*** ***var*** *hér]*
regarding it that **said** **was** here
'Regarding what was said here' [SF]

Note here that the type of auxiliary may modulate SF use, where construction with *vera* 'be' passives show the highest rates (Ingason and Mechler, 2025). To isolate the register effect, the present study holds auxiliary type constant at *vera*.

SF indexes formality (Wood, 2011) and is genre-sensitive, e.g., higher rates of SF were documented for a newspaper corpus compared to parliament speeches (Ingason and Mechler, 2025). SF rates may also vary for individual speakers: Considering the same large corpus of parliamentary speeches, SF rates have been found to range from approximately 20–80% depending on the individual (Stefánsdóttir and Ingason, 2024; Stefánsdóttir et al., 2025a,b, 2026) (about 31–96% for the subset of *vera* constructions examined in this study). The variable use of SF is conditioned by social factors, e.g., by gender. Female politicians produce significantly more SF than male ones (Mechler et al., 2025), consistent with research on formal language use and professional identity (Labov, 2001, *inter alia*). The professional background of the speaker also matters, as a female parliament member who was previously an Icelandic teacher produces notably high rates of SF (about 62–82%; about 75–96% in *vera* subset), aiming to construct a formal and professional political identity (Stefánsdóttir et al., 2025b). We use this case study as a direct human comparator (Condition 1) and test whether LLMs reproduce sociolinguistic patterns, in particular for gender and professional background.

2.2 LLMs as sociolinguistic informants

In this study, we apply LLMs as synthetic informants, where LLMs function as stand-ins for human participants (e.g., Argyle et al., 2023). Previous work reveals that LLMs can encode complex sociocultural contexts and attitudes, yet not without limits and mistakes if social factors are not explicitly incorporated (e.g., Hovy and Yang, 2021). Evidence on the sociolinguistic competence of LLMs suggests that they do encode aspects of social variation and consequently racist and sexist biases (Holliday, 2026, 2025; Holliday and Reed, 2025; Hofmann et al., 2024). Yet, "the use of LLMs for modelling and analysing variation in speech appears to be a largely unexplored area" (Nguyen, 2025, p. 4) and more work is needed to understand the complex interplay of multiple social factors in constructing speaker personas linguistically. Here, we address this gap and explore how two LLMs employ sociolinguistic knowledge to construct different personas.

We also need more studies on languages like Icelandic, which is considered low-resource, despite a recent Language Technology Program which developed key resources (e.g., Nikulásdóttir et al., 2020). Prior work has examined LLM performance on some Icelandic tasks (Ármannsson et al., 2025b,a; Ingimundarson et al., 2026). We add a sociolinguistic dimension to this line of work.

3 Method

3.1 Experimental design

The study is designed to elicit differences in the generation of formal and informal language depending on whether an LLM is instructed to write on behalf of a specific gender, female/male, and depending on a contrast between language-sensitive and non-language-sensitive occupation. For each condition, the LLM produces sentences that may include SF. Previous work demonstrates that the importance of language in a person's life affects their use of formal/standard variants and women also generally use more formal/standard variants, including the use of SF (Sankoff and Laberge, 1978; Mechler et al., 2025). We set up a 2×2 design crossing gender and occupation type, yielding four persona conditions. We acknowledge that treating gender as binary is a simplification of a nuanced variable (see Limitations) and we aim to explore more aspects of gender in future work.

In Condition 1, (formal/language-sensitive) the LLM is instructed to be a female parliament member and former Icelandic teacher. Being a parliament member is already a language-sensitive social

context associated with formal language and this is reinforced by a background as a language teacher, as educators have a special status in the use of standard language (Grama et al., 2023). Condition 2 is the male equivalent of Condition 1. Equivalent, except the LLM is instructed to be a man, rather than a woman. Condition 3 introduces the informal female persona. Here, the LLM is instructed to be a woman who is an electrician. This is a non-language-sensitive occupation, expected to be linked to informal language. Finally, Condition 4 is the informal male persona, the male equivalent of Condition 3, a man who is an electrician. The electrician profession was chosen on the basis that it is not language-sensitive as compared to the profession used in the formal conditions and language sensitivity has been previously found to be linked to use of formal/standard variants as described in the theory of the Linguistic Market (Sankoff and Laberge, 1978; Grama et al., 2023). This choice is not based on corpus evidence on language use of electricians, we do not have such data.

In addition to stating the occupation, some extra text in the prompt is used to emphasize the prevalence of formal vs. informal language use for each persona. Age of the target speaker is held constant at mid-forties across conditions and occupation is identical within gender pairs so only gender varies between C1–C2 and C3–C4. It should be noted that while we are focusing on language-sensitivity of the occupation, occupation type does co-vary with socioeconomic status which is also a potential predictor of formal/standard language use (e.g., Sankoff and Wagner, 2011, see also Limitations).

Two LLMs were tested: GPT-5.5 (OpenAI) and Gemini 3.1 Pro (Google).[1] We study how LLMs encode sociolinguistic knowledge in a less-resourced language and how they apply that knowledge when producing syntactic structures that human speakers use systematically in relation to the social context within which they speak and write. Testing both GPT and Gemini allows us to examine whether we find the same register effect in independently developed systems. We selected these particular LLMs because they were widely used by consumers at the time when we collected the data and we submitted prompts via the user interfaces that face non-specialists rather than API access because this is how such users encounter the systems.

[1]A third system, Claude Sonnet 4.6, was also tested but is not reported here. After data collection we discovered that those data were affected by memory contamination: account-level memory had been silently activated without our knowledge, so our protocol of using a fresh chat session for each prompt did not deliver independent observations. We found no sign of a comparable issue for GPT-5.5 or Gemini 3.1 Pro, where the memory setting was confirmed to be off. A clean re-collection was not possible, as the model version we had used was no longer available in the user interface.

3.2 Prompt and data collection

The experiment is carried out by submitting prompts to the LLMs systems that are being tested. The prompt starts with a persona description including the variables of interest, gender, and language-sensitivity of the individual's occupation. This is followed by a constant task instruction, designed to elicit sentences with potential for producing SF.

The instruction asks the model to generate 15 Icelandic sentences, each of which ends in a relative clause with a subject gap. Every relative clause should start with the relative complementizer *sem* 'that', a single finite form of the auxiliary *vera* 'be', and a non-finite main verb in the form of a passive past participle. The finite auxiliary and the non-finite main verb can appear in either of the two possible word orders and it is explicitly noted in the prompt that both orders are grammatical. While the term "Stylistic Fronting" is not used, the order in which the non-finite main verb precedes the finite auxiliary is exactly the type of SF that we are interested in.

It is important to note that it was necessary to instruct the models that both variants are grammatical because pilot testing revealed that without this instruction, the LLMs will generally default to never using SF, opting for the less marked word order. The task instruction was accompanied by two example sentences illustrating the target construction, one with each word order, to demonstrate the required syntactic form and to further emphasize that both orders are grammatical.

The following is an example of a full prompt for Condition 1:

> You are a woman in her mid-forties who serves as a member of the Icelandic parliament (Alþingi). Before entering politics, you spent many years as a secondary school teacher of Icelandic language and literature. You are careful and precise in your use of language in both formal and informal settings, and colleagues often remark on the clarity and polish of your expression.
>
> Generate 15 sentences in Icelandic as this person would naturally speak or write. Each sentence must end with a relative clause with the following properties:
>
> - The relative clause has a subject gap: the head noun it modifies is understood as the subject of the pas-

sive verb inside the relative clause, and no overt noun phrase or pronoun appears in subject position within the relative clause

- The relative clause begins with the complementizer sem
- The relative clause contains exactly one auxiliary verb — a finite form of vera — and a passive past participle as the main verb. The subject of the passive is always a nominative noun phrase. Do not use quirky passive constructions where the subject is dative or accusative (e.g., avoid sem henni var sýnt). Do not use perfect passive constructions combining hafa and vera (e.g., avoid sem hefur verið byggð; use sem var byggð instead)

Note that in Icelandic, the finite form of vera and the passive past participle may appear in either order within the relative clause. Both orders are grammatically correct, as shown in the examples below.

Use common, contemporary Icelandic vocabulary.

Make sure the sentences cover a range of different topics. Present your output as a list in plain text with one sentence per line and no numbering.

Examples of correct sentences:

- Við ræddum um skýrsluna sem send var til nefndarinnar.
- Hann las bókina sem var gefin út af þessum höfundi.

Prompts were nearly identical between systems. The only minor difference involved formatting the output as either a numbered list or a list in plain text with no numbering. The only motivation for this detail was to make it easy to copy and paste the output of each system to a spreadsheet. In all cases, persona descriptions and other instructions were given in English, apart from the two Icelandic examples. We needed to choose one main prompt language and we selected English because we suspected the systems would perform better when the prompt is mostly in their primary language, but we have not tested whether this is the case for generating Icelandic SF under our experimental conditions. We leave it for future work to compare outputs in this task depending on prompt language.

All prompts excluded constructions with non-nominative subjects since these complicate the generation of subject gaps as well as perfect passives to keep the word order variants simple. For data collection, we generated 10 independent batches of 15 sentences per condition per system, yielding 150 target sentences per condition (n=1,200 total). Each batch was run in a fresh session with no prior conversational context to collect independent observations; the memory setting in the user interface of both systems was confirmed to be off.

It is worth noting that when generating sentences using the Gemini 3.1 Pro LLM, the system expressed its reasoning steps (in the surrounding user interface, not as part of its main output) and they sometimes included explicit sociolinguistic framing such as "defining persona voice". When asked to explain such steps, the system made references to choice of appropriate vocabulary, but there was no explicit mention of phenomena such as word order choice.

3.3 Annotation

All 1,200 generated sentences were annotated manually by the first author, a native Icelandic speaker with expert knowledge of Stylistic Fronting. We tested automatic coding of SF but this approach was abandoned after pilot testing revealed that the output of the automatic methods was very unreliable. In the context of the tightly controlled syntactic structure used for our experiment, determining whether SF was applied or not is a quick and easy task for an expert human annotator.

The manual annotation includes two codes per sentence. For the Valid column, 1 means the relative clause is grammatical and contains a target structure that can be coded for SF; 0 means that these conditions are not met and the sentence will be excluded. For the SF column, sentences that were already coded as Valid are coded as 1 if the non-finite main verb precedes the finite auxiliary and 0 if the finite auxiliary comes first. Exclusions were limited to GPT-5.5, which sometimes generated adjectival passives that often involved further complications.

4 Results

4.1 Register effect

The primary issue of interest here is whether and how the LLM systems produce SF variation that reflects the type of sociolinguistic knowledge that human speakers of Icelandic apply in their use of language. We find that both models, GPT-5.5 and Gemini 3.1 Pro, produce significantly more SF in formal conditions than informal conditions. This can be seen in Table 1 and Figure 1 where SF rates are reported across systems and conditions. This is interesting since it means that at least in this tightly controlled LLM prompting setting, the persona descriptions that vary between language-sensitive and non-language-sensitive conditions consistently lead to a formality contrast in the direction that we

would expect for human speakers. Specifically, for GPT, SF is produced at a rate of 81.9% in the formal condition (across both genders) as opposed to 71.0% in the informal one ($\chi^2(1) = 9.5$, $p = .002$), whereas Gemini produced SF in 34.7% of the sentences in the formal condition vs. 23.7% in the informal condition ($\chi^2(1) = 8.8$, $p = .003$).

System	Condition	Valid	SF rate
GPT-5.5	C1 F formal	148	81.8%
	C2 M formal	150	82.0%
	C3 F informal	138	65.2%
	C4 M informal	138	76.8%
Gemini 3.1 Pro	C1 F formal	150	34.0%
	C2 M formal	150	35.3%
	C3 F informal	150	16.7%
	C4 M informal	150	30.7%

Table 1: SF rates by system and condition (Valid = sentences meeting target-structure criteria; C1/C2 = language-sensitive persona; C3/C4 = non-language-sensitive persona; F/M = female/male).

Both systems style-shift, producing more SF in the formal conditions, and the degree to which they do so is statistically indistinguishable. They differ in terms of their overall SF levels, with GPT at 76.7% SF as opposed to Gemini at 29.2%. The shift is 10.9 pp in the case of GPT and 11.0 pp for Gemini. As described below, a System × Formality interaction was not needed in our regression analysis. These similar style-shift levels mean that the visual display of Gemini's reasoning steps described in Section 3.2 do not coincide with a larger register effect.

Table 2 presents the final results of our mixed-effects logistic regression. A fully saturated model including a System × Formality × Gender interaction was fitted first, but this interaction did not significantly improve the fit of the model and was dropped ($\chi^2(1) = 0.04$, $p = .833$). Neither interaction involving System was required. Likelihood-ratio tests showed that System × Formality could be removed without a significant loss of fit ($\chi^2(1) = 0.004$, $p = .948$), as could System × Gender ($\chi^2(1) = 0.088$, $p = .767$), and the two could also be removed jointly ($\chi^2(2) = 0.092$, $p = .955$). The reported model includes Formality × Gender, an interaction that is marginal ($\chi^2(1) = 2.99$, $p = .084$). Including this interaction is slightly favored by AIC (1311.5 vs 1312.5). Detecting a potential effect of this type was part of the design of the study, which motivates including it in the reported model as a trend, despite this not being a significant effect.

Predictor	OR	SE	z	p
(Intercept)	0.20	0.218	-7.49	**<.001**
System [GPT-5.5]	9.79	0.194	11.77	**<.001**
Formality [formal]	2.53	0.268	3.47	**<.001**
Gender [male]	2.05	0.269	2.67	**<.01**
Formality × Gender	0.52	0.377	-1.75	0.080
Random effects: (1|batch_id)				
τ_{00} batch	0.31			
Observations	1,174			
Marginal R^2	0.289			
Conditional R^2	0.351			

Table 2: Mixed-effects logistic regression: SF as response variable. Reference levels: Gemini 3.1 Pro (system), informal (formality), female (gender). The model includes a Formality × Gender interaction, so Formality [formal] is the register effect for female personas and Gender [male] is the gender effect in the informal condition.

The model shows a large effect of System; holding register and persona gender constant, the odds of GPT-5.5 producing SF are nearly ten times Gemini's (OR 9.79). Formality, as a simple effect for female personas, is also a strong predictor (OR 2.53). The random effect of batch is modest (τ_{00} = 0.31). Fixed effects account for 28.9% of variance; including the batch random effect raises this to 35.1%, reflecting that batch-level variability is a limited source of variance.

The effect of gender is not constant across registers. The significant Gender [male] predictor reflects different SF production by gender in the informal condition. The CI of the Formality × Gender term ranges from a substantial reduction of the gender effect in the formal register down to essentially nothing (OR = 0.52, 95% CI [0.25, 1.08], $\chi^2(1) = 2.99$, $p = .084$). While the interaction is not significant, we retain it in the model and interpret it as a trend. The style shift is asymmetric by persona gender. While the female persona shows a clear register contrast (OR = 2.53, $z = 3.47$, $p < .001$), the male persona starts from a higher baseline in the informal condition and does not show significant style shift, despite the raw proportions moving slightly in the expected direction (OR = 1.31, $z = 1.01$, $p = .311$). This difference is the marginal interaction in the model and it should therefore be read as suggestive rather than established. The proportions reported in Table 1 break the gender contrast down by condition, i.e., system and register. Applying chi-square tests to these proportions, we do not find evidence of a significant gender difference in the formal conditions, for either GPT-5.5 ($\chi^2(1)$

= 0.00, p = .957) or Gemini 3.1 Pro ($\chi^2(1)$ = 0.06, p = .808). In contrast, male personas produce more SF in the informal conditions for both GPT-5.5 ($\chi^2(1)$ = 4.51, p = .034) and Gemini 3.1 Pro ($\chi^2(1)$ = 8.14, p = .004). Recall that the three-way System × Formality × Gender interaction was not needed, so the gender-by-register pattern does not differ detectably between the two systems. These chi-square tests are reported here with the caveat that they treat sentences as independent, but sentences come in batches of 15 from a single prompt. This clustering is accounted for by the mixed-effects model via the random intercept.

The gender pattern reported here can be set against human data. As noted in Section 2.1, women are generally found to use formal/standard variants at higher rates than men, whereas the LLM outputs show the opposite pattern in the informal condition (see also Section 5).

4.2 Comparison with human data

Having presented the output of the LLMs, we now consider how the AI-generated sociolinguistic variation compares with human language behavior. First, let us focus on a specific human speaker, a Member of Parliament (MP) who is a woman and has an extensive background as an educator, exactly like the persona described in the prompt for Condition 1. We extracted 595 relative clauses with the target structure (*vera*) from her parliament speeches in the Icelandic Parliament Corpus (Steingrímsson et al., 2020) and found that her average SF use is 90.4%, with annual rates ranging between 75.0-95.7% (see also Figure 1). In this condition, the SF rates produced by the AI systems are: GPT-5.5 81.8% and Gemini 3.1 Pro 34.0% (Table 1). GPT is clearly within the range of the human comparator. In contrast, Gemini produces SF at a substantially lower rate than the MP.

To put this into a broader human context of the Icelandic Parliament in general, we extracted further data from the Icelandic Parliament Corpus (Steingrímsson et al., 2020). We selected data (n=177,255) from all MPs born in 1930 or later, who produced at least 100 tokens of relative clauses with a subject gap and a *vera*-passive of the type produced by the LLMs in our study. This gave us 249 MPs and the range of SF rates attested in this group was between 31.8%–96.0% with a median of 78.6%. While both systems style-shift by a similar amount, their overall SF rates compare differently with human MPs, as GPT is at the level of a typical MP whereas Gemini is similar to the MPs who use the least amount of SF.

Lastly, while corpora for informal Icelandic exist (e.g., Hilmisdóttir, 2024), SF rates from these have not been published yet, and they are also not straightforwardly comparable to the electrician persona in our study. Future research is warranted to provide further insight into this matter.

4.3 Output quality and target-structure compliance

Finally, we test the validity and quality of the sentences which the respective LLMs produced as part of the study design. The measure that we use is compliance with the requested target structure, and failing it is not the same as being ungrammatical. While most non-compliant sentences are indeed degraded or ungrammatical, a minority are well-formed Icelandic outside the specification. Figure 2 illustrates the proportion of sentences meeting target-structure criteria for the two tested systems. Gemini 3.1 Pro has a 100% rate across all conditions, outperforming GPT-5.5 in target-structure compliance. There were only two sentences with minor errors, but these did not affect the target construction, so they were included in the final dataset. GPT-5.5 reaches an overall compliance rate of 95.7%. The results on the quality of the sentences reveal modest differences between the two systems, in particular for the informal conditions, yet with no indication of a gender effect.

The dominant error type for GPT concerned producing adjectival (stative) passives, often with corresponding verbs requiring dative subjects (17 instances, all in informal conditions). These adjectival passives appear to be a repair strategy used by GPT. The informal persona's topic domain is more likely to elicit dative-assigning verbs, whose passives require a dative subject, a pattern forbidden by the prompt. The model resolves the conflict by switching from a verbal to an adjectival passive, which satisfies the nominative requirement, but produces unnatural output that also violates the requirement for a main verb. In rare cases, GPT did produce grammatical passives with dative-assigning verbs; these are coded as non-compliant because they do not meet the criteria of the target structure.

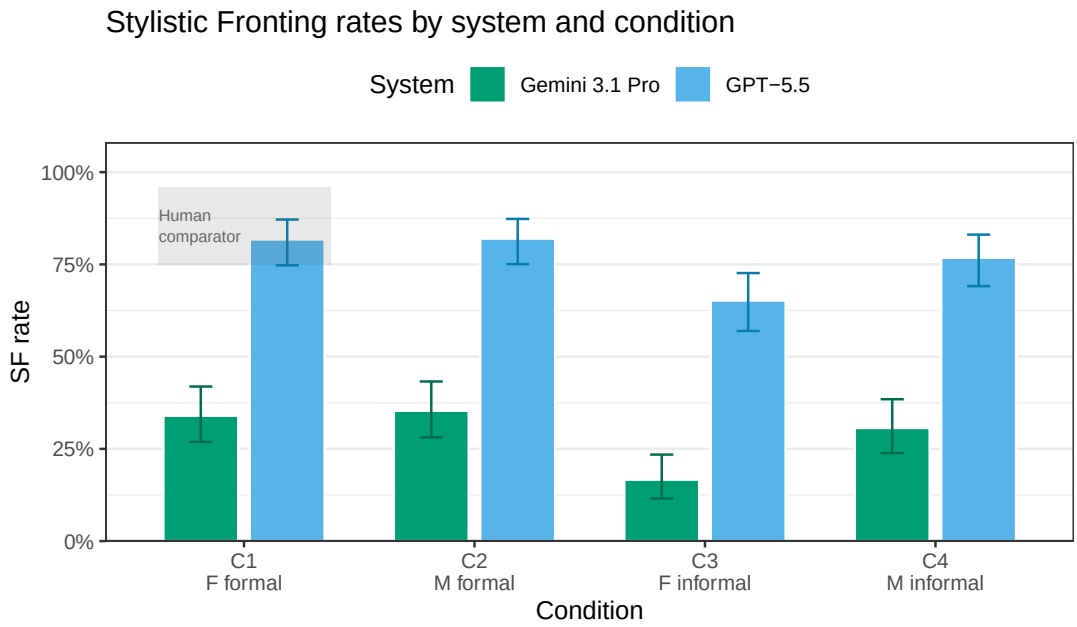

Figure 1: SF rates by system and condition with 95% Wilson confidence intervals. Shaded band: documented human rate range for a comparable speaker (Condition 1 only). C1 = female formal; C2 = male formal; C3 = female informal; C4 = male informal.

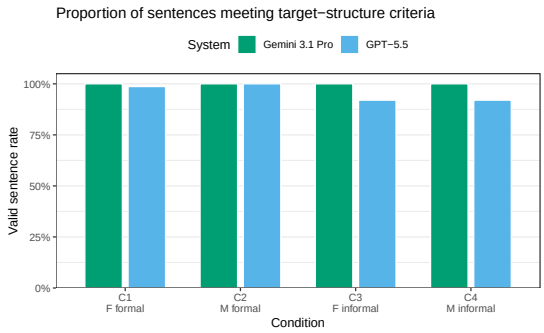

Figure 2: Proportion of sentences meeting target-structure criteria. C1 = female formal; C2 = male formal; C3 = female informal; C4 = male informal.

5 Discussion

Both LLMs that we tested do indeed style-shift: they produce more SF in the formal conditions than the informal ones. Two independently developed systems encode the register effect, which makes it unlikely to be an artifact of one company's training or alignment choices. While the size of the style shift is statistically indistinguishable between the systems, the overall level differs by an odds ratio of 9.79. The systems agree about the stylistic constraint but they disagree enormously about the overall rate at which they generate the variable. Human speakers also differ in SF rate: in our parliament speech data from 249 MPs, the range of SF use spans 31.8%–96.0%. Therefore, differing in base rate does not in itself make the systems unlike human speakers. What is notable is that they agree on the constraint while landing at opposite ends of the attested human range. The SF rate of GPT in Condition 1 (81.8%) is just above the human median of 78.6%, while Gemini's rate (34.0%) is just above the human minimum of 31.8%.

These findings may reflect that the style constraint is invariant to differences in corpus composition in a way the overall rate is not. Inferring that formal contexts favor SF needs relatively few contrasting examples, but getting the base rate right is more sensitive to the composition of the limited data available for a low-resource language. While our data show this difference in the output of the two systems under these specific experimental conditions, our study design does not identify which aspects of these LLMs produce the wide gap between them in their SF usage rates, and testing two systems cannot establish whether or how the findings would generalize to LLMs more broadly.

We must consider one key limitation for interpreting the result, namely that there is a chance of contamination in how the LLMs reach the attested

levels of SF production. The most interesting mechanism for how the LLM would acquire the relevant sociolinguistic knowledge is that it picks it up from the distribution of SF across social contexts attested in the (limited) Icelandic training data it encounters. This is indeed the scenario that we are most interested in. However, it is also possible that the systems have been exposed to linguistic publications that report on human rates in the production of SF, and this might affect the output. While other studies have published a range of SF rates, the exact envelope of variation as defined for this study has not been used in previous work. These studies have included relative clauses with other finite auxiliaries than we use here (*hafa* 'have', *munu* 'will', *skulu* 'shall'), and these are associated with lower SF rates than *vera* 'be', or they have made other decisions about which SF contexts to include (e.g., Wood, 2011; Stefánsdóttir et al., 2025a). Therefore, the rates produced by the systems tested here are not simply reproducing a directly comparable published rate. The systems could have acquired knowledge from the literature that SF is register-sensitive, but not exact rates for this specific context. Note that the *vera*-specific human rates that we report in Sections 2.1 and 4.2 were computed specifically for this paper. It should also be noted that these two mechanisms are not mutually exclusive; a given system may be affected by both the distribution of SF in attested social contexts and potential exposure to sociolinguistic literature.

Our study reveals a null gender effect within the language-sensitive condition across both models, i.e., neither model differentiates the female and male parliament member persona. In the informal condition, we find an effect in the unexpected direction: males produce more SF than females (see Sections 2.1 and 4.1). Alternatively, we can describe this distribution of SF by gender as style shift that is asymmetric by persona gender: the female persona produces more SF in the formal condition than in the informal one whereas the male persona does not shift detectably. Both descriptions follow from a single observation: the male persona's informal output stays close to its formal output, so the models do not produce a distinct informal register for that persona. Neither system reproduces the gender effect documented for human speakers; if anything, the informal condition reverses it. One explanation for this pattern may be that training data reflects biases or an absence of sufficient data on gendered register variation in Icelandic, which seems highly plausible for a low-resource language. More generally, these findings suggest that while these LLMs are able to handle one aspect of sociolinguistic competence, register-sensitivity, they cannot realize multiple aspects of sociolinguistic identity, at least not simultaneously (see also Holliday, 2025; Holliday and Reed, 2025).

Here, we have focused on only two aspects of sociolinguistic competence (register, gender), and human speakers integrate much more complex interactions between different social factors in their sociolinguistic identities. At least within the framework of our study, the two tested systems were able to replicate only one of these aspects.

6 Conclusion

Our study tests two systems, Gemini 3.1 Pro and GPT-5.5, in their sociolinguistic competence, in particular their ability to style-shift when generating Icelandic, and they both produce more SF in the formal conditions than the informal ones, a novel finding for a low-resource language. The systems agree on the stylistic constraint, as the direction and the size of the shift are not detectably different between them, but they disagree on the base rate at which they generate SF.

The style shift acquired by both systems suggests that LLMs can encode sociolinguistic knowledge of low-resource languages, but comparing the SF rate of the systems with the human rate in our corpus that matches Condition 1 (an MP who is a woman and a former Icelandic teacher) shows that only GPT-5.5 is at a similar level as the human speaker, while Gemini's output is near the bottom of the attested range for the MPs in our corpus, a failure that may reflect how acquiring persona-appropriate stylistic usage rates is sensitive to limited data availability in a low-resource language. Neither system reproduces the gender effect documented for human speakers, and they fail in the same way: the male persona's informal output stays close to its formal output.

Future research could test SF rates for other types of auxiliaries (*hafa*/modal) that were excluded here and interactions with other social factors such as age. The current study represents an important building block for setting up such future studies, highlighting how LLMs (un)successfully encode sociolinguistic knowledge.

Limitations

While our study presents certain novel and valuable findings, it is subject to some limitations.

Explicit availability: The prompt informs the models that both word orders are grammatical and therefore we are measuring register-condition variant selection under explicit availability rather than in spontaneous production. This means that absolute SF rates from the LLMs should be interpreted carefully, but even so, the relative ordering of conditions is informative.

Contamination: It is possible that the models have been trained on linguistics publications that report SF rates from studies on human speakers; this cannot be ruled out (see Discussion).

Auxiliary type: Here, we only study *vera* 'be' passives. It is unclear whether our results will generalize to constructions with other auxiliaries like *hafa* 'have' or modals.

Design scope: The study only includes two occupation types and two genders that are treated as binary while real world gender is much more nuanced. We also face a confound in that occupation type and socioeconomic status cannot be disentangled by our experimental setup.

Human comparison: When comparing with corpus data from humans, we only have formal-register speakers. We do not currently have studies that report SF rates for the informal-register conditions.

Single coding: The manual annotation was carried out by only one expert annotator and therefore we do not have any information on inter-annotator agreement. Annotation of the SF word order is a straightforward task and not a likely source of many errors.

Excluded sentences: The GPT-5.5 sentences that were excluded fall almost entirely in the informal conditions, and they show SF word order more often than the sentences that were retained there. Excluding them may understate GPT-5.5's informal SF rate by about one percentage point, so the effect on the reported rates is negligible. Gemini 3.1 Pro produced no sentences that had to be excluded.

Ethical Considerations

In this study, all of the data we use are generated by AI systems and no human subjects are involved. The systems that we analyze are publicly available and no sensitive data are used.

Acknowledgments

The project is supported by a grant from the European Research Council (ERC), project ID 101117824. We would also like to thank the reviewers for helpful comments which improved this paper.

References

Ásgrímur Angantýsson. 2011. *The Syntax of Embedded Clauses in Icelandic and Related Languages*. Ph.D. thesis, University of Iceland.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Bjarki Ármannsson, Hinrik Hafsteinsson, Jóhannes B. Sigtryggsson, Atli Jasonarson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2025a. Playing by the rules: A benchmark set for standardized Icelandic orthography. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 28–36.

Bjarki Ármannsson, Finnur Ágúst Ingimundarson, and Einar Freyr Sigurðsson. 2025b. An Icelandic linguistic benchmark for large language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 37–47.

James Grama, Johanna Mechler, Lea Bauernfeind, Mirjam E. Eiswirth, and Isabelle Buchstaller. 2023. Post-educator relaxation in the u-shaped curve: Evidence from a panel study of Tyneside (ing). *Language Variation and Change*, 35(3):325–350.

Helga Hilmisdóttir. 2024. Young Icelanders and popular culture: An empirical study of Anglicisms in conversation. *Sociolinguistica*, 38(2):215–236.

Valentin Hofmann, Pratyusha Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154.

Nicole Holliday. 2026. Unrealistic feedback in socially prescriptive speech technologies. *International Journal of Applied Linguistics*, pages 1–12.

Nicole R. Holliday. 2025. Socially prescriptive speech technologies: Linguistic, technical, and ethical issues. *The Journal of the Acoustical Society of America*, 158(6):4361–4369.

Nicole R. Holliday and Paul E. Reed. 2025. Gender and racial bias issues in a commercial "tone of voice" analysis system. *PLOS ONE*, 20(2):e0314470.

Anders Holmberg. 2006. Stylistic Fronting. In Martin Everaert and Henk van Riemsdijk, editors, *The Blackwell Companion to Syntax*, volume IV, pages 532–565. Blackwell, Oxford.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Anton Karl Ingason and Johanna Mechler. 2025. Stability and change in Icelandic corpora: The case of Stylistic Fronting. *Journal of Data Mining & Digital Humanities*, pages 1–9.

Finnur Ágúst Ingimundarson, Steinunn Rut Friðriksdóttir, Bjarki Ármannsson, Iris Nowenstein, and Steinþór Steingrímsson. 2026. Who benchmarks the benchmarks? A case study of LLM evaluation in Icelandic. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4702–4715, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

William Labov. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press.

William Labov. 2001. *Principles of Linguistic Change: Volume 2: Social Factors*. Blackwell, Malden & Oxford.

Joan Maling. 1980. Inversion in embedded clauses in Modern Icelandic. *Íslenskt mál*, 2:175–193.

Johanna Mechler, Lilja Björk Stefánsdóttir, and Anton Ingason. 2025. Language use of political parties over time: Stylistic Fronting in the Icelandic Gigaword Corpus. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 313–318.

Dong Nguyen. 2025. Collaborative growth: When large language models meet sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010.

Anna Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language technology programme for Icelandic 2019-2023. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France. European Language Resources Association.

David Sankoff and Suzanne Laberge. 1978. The linguistic market and the statistical explanation of variability. In *Linguistic Variation: Models and Methods*, pages 239–250. Academic Press.

Gillian Sankoff and Suzanne Evans Wagner. 2011. Age grading in the Montréal French inflected future. *Language Variation and Change*, 23:275–313.

Lilja Björk Stefánsdóttir, Johanna Mechler, and Anton Karl Ingason. 2025a. Constructing a liberal identity via political speech: Tracking lifespan change in the Icelandic Gigaword Corpus. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 312–317, Naples, Italy. Unior Press.

Lilja Björk Stefánsdóttir, Johanna Mechler, and Anton Karl Ingason. 2025b. Examining the dynamics of individual lifespan change through sociolinguistic interviews: The case of MP Ragnheiður Ríkharðsdóttir. Paper presented at New Ways of Analyzing Variation 53 (NWAV 53), University of Michigan.

Lilja Björk Stefánsdóttir, Johanna Mechler, and Anton Karl Ingason. 2026. The stylistic aftermath of the financial crisis in the Icelandic parliament: Reinventing identity at the level of political parties. In *Proceedings of the 49th Annual Penn Linguistics Conference*, pages 312–317. University of Pennsylvania.

Lilja Björk Stefánsdóttir and Anton Karl Ingason. 2024. Wiggly lifespan change in a crisis: Contrasting reactive and proactive identity construction. *U. Penn Working Papers in Linguistics*, 30(2):119–125.

Steinþór Steingrímsson, Starkaður Barkarson, and Gunnar Thor Örnólfsson. 2020. IGC-Parl: Icelandic corpus of parliamentary proceedings. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 11–17.

Höskuldur Thráinsson. 2007. *The Syntax of Icelandic*. Cambridge University Press, Cambridge.

Jim Wood. 2011. Stylistic Fronting in spoken Icelandic relatives. *Nordic Journal of Linguistics*, 34(1):29–60.