PRESERVING LINGUISTIC DIVERSITY IN THE AI ERA: ANALYZING ENGLISH INTRUSION IN DANISH LLM OUTPUTS

TECHNICAL REPORT

Mattis Kragh (s246222)
BSc programme Artificial Intelligence and Data
Technical University of Denmark
Copenhagen, Denmark

Ro Nanak Prasad Lacoul (s246152)
BSc programme Artificial Intelligence and Data
Technical University of Denmark
Copenhagen, Denmark

Antonijs Bolsakovs (s225124)
BSc programme General Engineering
Technical University of Denmark
Copenhagen, Denmark

Christopher Alexander Holm Kjær (s234744)
BSc programme Artificial Intelligence and Data
Technical University of Denmark
Copenhagen, Denmark

June 18, 2025

ABSTRACT

Large Language Models (LLMs) are central to digital interactions, despite English remaining dominant in their training and output, raising concerns for smaller language communities. This report investigates unprompted English intrusion in Danish-language outputs from Google Gemini 1.5 Flash and 2.0 Flash. We hypothesize that newer versions, especially Gemini 2.0 Flash, will show a statistically significant increase in English word intrusion, particularly in longer, more complex responses and specific prompt categories. We analyzed generated responses across academic, creative, riddle, and FAQ prompts. Our methods included quantifying English content, Association Rule Mining for Danish word triggers, ANOVA for variable effects, Pearson correlation for response length relationships, and Mann-Whitney U tests. Findings indicate Gemini 1.5 Flash maintains stable Danish output regardless of length, while Gemini 2.0 Flash exhibits significantly stronger positive correlations between response length and English word frequency, notably in STEM, FAQ, and general categories. These results suggest AI systems could inadvertently normalize English in Danish communication, underscoring important implications for linguistic autonomy and digital language diversity.

Keywords Generative Artificial Intelligence · Large Language Models · Danish Language · English Language · Linguistic Diversity · Sociolinguistics

Contents

1 Introduction

In recent years, large language models (LLMs) have become central to how people interact with digital systems, performing tasks like writing, translating, and answering questions. As LLMs become more integrated into daily life, their influence on communication and education grows.

A key feature of modern LLMs, such as Google Gemini, is their multilingual capability. While trained on data from many languages, English remains dominant—both in the training data and in model behavior. This reflects not only technical bias but also the global influence of English [Smith & Jones (2025)].

This imbalance becomes especially clear with smaller languages like Danish. Even when prompts are fully in Danish, LLM outputs often include unexpected English words or phrases. This raises concerns about how AI might shape language use and influence linguistic norms.

From a sociolinguistic perspective, repeated exposure to English terms in AI-generated Danish—especially in educational or formal contexts—could contribute to language shift or erosion over time [Gottlieb (2004)].

This report investigates how Gemini 1.5 Flash and 2.0 Flash respond to Danish-language prompts. We analyze the frequency and context of English elements in their outputs, aiming to shed light on language dominance in AI and its implications for linguistic diversity.

2 Problem and Motivation

Large scale language models such as Google Gemini [Google (2025)] promise seamless multilingual interaction yet English still dominates both their training data and many of their generated responses. This imbalance has broader implications for smaller languages like Danish, frequent and unintentional exposure to English words in ai generated content can gradually make anglicisms seem normal and lead to the decline of Danish vocabulary in in fields such as technology or education.

This project investigates how consistently Gemini 1.5 Flash and 2.0 Flash maintain Danish in response to Danish prompts, and what factors lead to English intrusions. Technically, we examine model consistency and statistical patterns behind code-mixing[1]; culturally, we consider whether LLM outputs might reinforce a shift toward English in Danish digital spaces

We hypothesize that newer versions of large language models, exemplified in Google Gemini 1.5 Flash and 2.0 Flash, will exhibit a statistically significant increase in unprompted English word intrusion in Danish language outputs compared to Gemini 1.5 Flash, particularly in longer and more complex responses and across specific prompt categories.

To guide our analysis, we pose five subquestions:

1. **Prevalence of mixing**: What proportion of otherwise Danish prompts elicit responses containing at least one English element?
2. **Topical sensitivity**: Across prompt categories (RIDDLES, FAQ, CREATIVE, STEM, NON-STEM, and INTER-disciplinary), which show the highest average share of English tokens?
3. **Response length effects**: Is there a statistically significant correlation between response length and the number of English words in the response?
4. **Lexical triggers**: Which specific Danish words or phrases are most strongly associated with subsequent English borrowings, as revealed by association rule mining?
5. **Nature of borrowings**: What kinds of English words appear (technical terms, slang, general vocabulary, brand names), and how do their distributions differ across prompt categories?

[1] A linguistic phenomenon several different languages are used together in a single conversation or exchange

The next section describes our data collection process and analytical methods in detail.

3 Methodology

This study used Google's Gemini 1.5 Flash and 2.0 Flash models to examine English word intrusion in Danish-language outputs. Due to English's dominance in training data, its influence on large language models like Gemini is notable. Although Gemini 1.5 Flash was deprecated on September 24, 2024, it remains accessible via API, allowing comparison with the still-active 2.0 Flash (available until at least February 5, 2026). This enabled analysis of potential shifts in English word usage across model versions, despite their close release dates.

3.1 Association Rule Mining

Association Rule Mining (ARM) was employed to discover if specific Danish words in a prompt correlate with the appearance of unprompted English words in the model's answer. The goal was to find patterns of the form:

$$\{\text{Words in Prompt}\} \implies \{\text{English in Answer}\}$$

We used the FP-Growth algorithm with a minimum support of 10% and a minimum confidence of 50%. Rules containing short, high-frequency words (≤ 3 characters) were excluded to focus on more substantive insights into English intrusion triggers. Experimentation with a 2-character cutoff did not yield additional useful results. This 10% threshold was chosen as a pragmatic trade-off between available computational resources (RAM) and the desire to extract a useful number of meaningful rules To ensure the discovered rules were meaningful and not just statistical noise from common grammar, we also applied the filter of a minimum confidence threshold of 50%

This methodology allows us to isolate substantive Danish prompt words that are most predictive of English word intrusion.

3.2 Correlation

To test whether longer prompts lead to more English interference in ai generated Danish responses, we performed a correlation analysis. Specifically we calculated the Pearson correlation coefficient (r) between the total number of words in a response and the number of English words it contained.

This analysis was carried out for both Gemini 1.5 Flash and 2.0 Flash across six prompt categories. We used `scipy.stats.pearsonr` to compute the r and the corresponding p value for each category. Statistically significant results ($p < 0.05$) indicate that English word usage increases in tandem with prompt length.

The outputs of this correlation analysis are further discussed and visualized in Section 6.2.

3.3 ANOVA

A Two-Way Analysis of Variance (ANOVA) was conducted to examine how two main factors — LLM version and prompt type — influence the proportion of English in generated responses. Two-Way ANOVA extends the basic ANOVA by analyzing the effects of two categorical variables and their interaction on a continuous outcome. It assumes independent observations, approximately normal residuals, and equal variances across groups.

We chose this method because our study involves two categorical factors: LLM version (Gemini 1.5 Flash and 2.0 Flash) and prompt type (e.g., Riddles, FAQ, Creative, STEM, Non-STEM, Interdisciplinary). Two-Way ANOVA allows us to assess their individual effects as well as how they interact.

This approach offers a comprehensive statistical framework, making it possible to evaluate both main and interaction effects in a single test — rather than relying on separate comparisons.

The results of this ANOVA are presented and discussed in the Analysis section (see Table 5).

3.4 Shapiro-Wilk

The **Shapiro-Wilk test** is used to assess whether a sample $\{x_1, x_2, \ldots, x_n\}$ comes from a normally distributed population. It tests the null hypothesis H_0: "the data are drawn from a normal distribution."

The test statistic is defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where:

- $x_{(i)}$ are the ordered sample values,
- $\bar{x}$ is the sample mean,
- a_i are constants derived from the expected values of order statistics of a standard normal distribution.

A value of W close to 1 suggests normality. A small p-value (typically < 0.05) leads to rejecting H_0, indicating that the data significantly deviate from a normal distribution. [Shapiro & Wilk (1965)]

3.5 Categorization of English Word Intrusion

A qualitative analysis was conducted to understand the nature of English words in Danish responses. Potential English words were first extracted by identifying words not on the official Danish wordlist (Det Danske Sprog- og Litteraturselskab) and then confirmed against the NLTK English corpus. An automated script using the Gemini 2.0 Flash API then classified each unique English word into one of five categories:

- **Technical Terms:** Specialized vocabulary (e.g., "algorithm", "interface").
- **General Vocabulary:** Common English words despite Danish equivalents (e.g., "answer", "important").
- **Brand Names:** Proper nouns for companies/products (e.g., "Google", "iPhone").
- **Slang:** Informal expressions (e.g., "cool", "awesome").
- **Other:** Words not fitting above, including misidentified words or non-brand proper nouns.

This LLM-based categorization was chosen for efficiency given the large dataset, though it introduces potential bias, as Gemini 2.0 Flash was used to categorize its own output without sentence context. While this limits objectivity and complicates error estimation, a small subset of samples was manually reviewed to informally verify overall categorization trends.

3.6 Mann-Whitney U Test

For the Mann-Whitney U test the null hypothesis is that the two populations are equal and the alternative hypothesis is that the two populations are not equal. It works under the assumptions that the variables are continuous in the two groups, the data must be non-normal and in similar shape. They must also be independent and use sufficient sample size. The result of the test will tell us if the two groups are significantly different, but to tell which group is an improvement over another, we must look at factors like means. [Technology Networks (2023)]

4 Data Collection

To generate responses for evaluation, we developed a dataset using prompts processed through the API for Google's Gemini 1.5 Flash and Gemini 2.0 Flash models.

For estimating a sample size we utilized the sample size for a proportion for a 95% confidence interval, as our results are given in a ratio between English and total words. Since we do not know this proportion we use $p = 0.5$ for the maximum sample size (even though this is an unlikely proportion), with a margin of error of 0.075 and get

$$n = \frac{Z^2 p(1-p)}{E^2} = \frac{1.96^2 \cdot 0.5^2}{0.075^2} \approx 171.$$

However, we decided to use all the daily API tokens we had at our disposal over a period of two days for three different users - see the specific amount of samples for a given prompt type in the table below. This means that we have a sufficient amount of data. These prompts spanned a variety of domains and styles, including creative, academic, and interdisciplinary topics.

The prompts were designed to broadly reflect how an average user might interact with LLMs. All prompts were crafted in Danish, with any exclusively English words replaced by their Danish equivalents to ensure pure Danish input. We acknowledge this as a limitation: without dedicated research into typical prompting behaviors across various user demographics or professional fields, our prompt design has an inherent bias towards the prompt design of university students.
All the data is independent of each other, since when we connect to the API, we are each time sending a new, independent request to the Gemini model. There is no memory retaining pertaining to that instances' prompt.

Riddles	**FAQ**	**Creative Prompts**	**Academic Questions**
Questions meant to require thinking to solve, ranging from nonsensical questions to riddles.	Frequently asked questions asked by users, these are exemplified in arbitrarily chosen context	Prompts where the LLM is asked to, for example, write something or do something with a great degree of freedom	Academic questions split into STEM, Non-STEM, and mixed fields
$n_{1.5} = 360$	$n_{1.5} = 360$	$n_{1.5} = 422$	$n_{1.5} = 360$
$n_{2.0} = 553$	$n_{2.0} = 570$	$n_{2.0} = 604$	$n_{2.0} = 559$

Table 1: Prompt Sections

n_x describes the number of answers under a given category, for a specific model. There was created 30 prompts for each section, and since generative AI, gives different responses most of the time, we ran each question several times to potentially test the reliability of the occurrence of English words in each prompt.

The script `reading_prompts.py` was used to automate the data generation process. It iterates through all prompt files, splits the content into individual prompt blocks, and sends each to the Gemini API using the `genai` client. To manage rate limits, a brief delay was inserted between requests.

The 30 unique prompts classified as 'Academic Questions' were further sub-categorized into STEM, Non-STEM, and Interdisciplinary topics. Due to API limitations and the iterative nature of prompt design, these 30 prompts were distributed such that the 'Interdisciplinary' sub-category contained the fewest unique prompts among these three.

Each model response was stored as a separate `.txt` file in the `output` directory for the given model. Filenames include metadata such as the source file, timestamp, and prompt index to ensure traceability.

5 Results

Source	sum_sq	df	F	PR(>F)
C(llm_version)	0.001967	1.0	0.630095	$4.273702e-01$
C(prompt_type)	0.149429	5.0	9.573535	$4.315340e-09$
C(llm_version):C(prompt_type)	0.042836	5.0	2.744381	$1.761402e-02$
Residual	11.787579	3776.0	NaN	NaN

Table 2: ANOVA Results Summary

The above table is the results of performing a two-way ANOVA test on the proportion of English words, from both llm models and all the prompt data gathered.
The means and 95% confidence intervals for the proportions of English words per category of prompts are given in the table below.

Metric	Riddles	FAQ	Creative Prompts	STEM	Non-STEM	Inter-disciplinary	Academic Total
Mean 1.5	0.616%	2.655%	0.780%	0.277%	0.147%	0.445%	0.234%
Confidence interval 1.5	[0.043%, 1.188%]	[1.565%, 3.745%]	[0.490%, 1.069%]	[0.191%, 0.363%]	[0.098%, 0.195%]	[0.252%, 0.638%]	[0.186%, 0.283%]
Mean 2.0	0.431%	1.429%	1.434%	0.307%	0.287%	0.346%	0.302%
Confidence interval 2.0	[0.232%, 0.629%]	[0.829%, 2.030%]	[0.878%, 1.989%]	[0.245%, 0.370%]	[0.233%, 0.341%]	[0.231%, 0.461%]	[0.263%, 0.341%]

Table 3: Means and 95% confidence intervals for proportions of English words for both Gemini 1.5 Flash and Gemini 2.0 Flash.

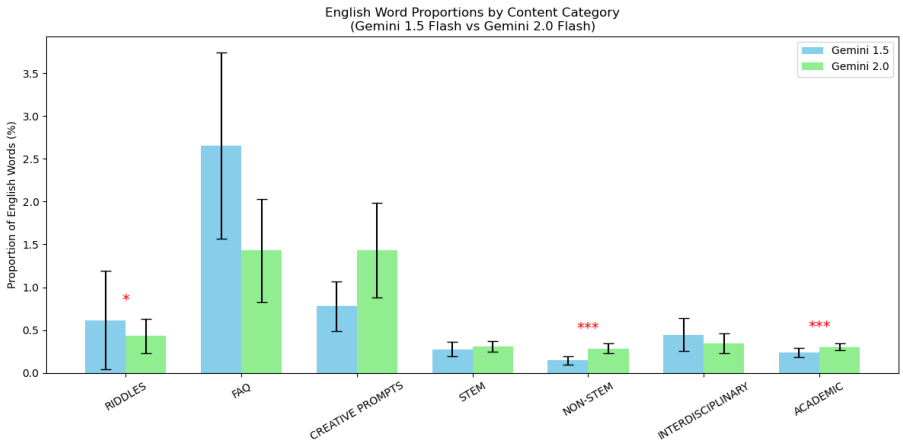

Figure 1: Comparison of means for English word proportions. * means significant difference after applying Mann-Whitney U test and *** means highly significant difference. See appendix for p-values from Mann-Whitney U test.

Category	Gemini 1.5		Gemini 2.0		Δr
	r	p	r	p	(2.0 – 1.5)
STEM	0.0026	0.9757	0.3194	**0.0000**	+0.317
FAQ	−0.0009	0.9858	0.0970	**0.0202**	+0.098
RIDDLES	0.1749	**0.0008**	0.0675	0.1119	−0.107
CREATIVE PROMPTS	−0.0136	0.7798	−0.0425	0.2961	−0.029
MULTIPLE	0.1429	0.3170	0.0294	0.7971	−0.114
NON	−0.0026	0.9723	0.2449	**0.0000**	+0.247

Table 4: Pearson correlation (r) between response length and English word count for Gemini 1.5 Flash and 2.0 Flash.
Bold p-values denote statistical significance at $\alpha = 0.05$.

Category	Gemini 1.5 (Count)	Gemini 1.5 (%)	Gemini 2.0 (Count)	Gemini 2.0 (%)
General Vocabulary	822	78.8%	1490	76.5%
Other	111	10.6%	254	13.0%
Technical Terms	71	6.8%	155	8.0%
Slang	30	2.9%	40	2.1%
Brand Names	9	0.9%	10	0.5%
Total	**1043**	**100.0%**	**1949**	**100.1%**

Table 5: Categorical distribution of unique English words for Gemini 1.5 Flash and 2.0 Flash. Percentage rounding may lead to slight total deviation.

6 Analysis

6.1 ANOVA

Test	Statistic (W)	p-value
Shapiro-Wilk	0.2016	1.621×10^{-83}

Table 6: Results of the Shapiro-Wilk Test for Normality of Residuals

Given the information from the two-way ANOVA table, we can see that there is a very high statistically significant effect of the prompt type, which means that the usage of English words varies significantly between different prompt types (p-value < 0.05). In the same way, we have also found a statistically significant interaction effect between prompt type and LLM version. This suggests that even if there is no significant difference on effect between models, it means that the difference between LLM versions regarding English word usage is not consistent across all prompt types, rather, the effect of the LLM version depends on the specific prompt type. Given the significant interaction effect, the individual main effects should be interpreted with caution, as the interaction indicates a more nuanced relationship.

To be certain of the accuracy of the ANOVA test, we performed a Shapiro-Wilk test, to see if the assumption of normality was true for the residuals. The conclusion of this test was that they were not normally distributed.

While ANOVA is generally robust to violations of normality with large sample sizes, the severity of this violation suggests that the p-values should be interpreted with caution. Though there is a slight risk that the significance of prompt types influence is not significant, the results from the ANOVA overlap with our other findings, so even though the residuals are not normally distributed, it's most likely still true that it is significant. The same is not necessarily true for the other effects, given their p-values being too close to the limit of significance, but there is no certain answer to be gained from the ANOVA analysis.

6.2 Answer Length vs English word usage

To investigate whether longer answers result in greater English interference in Gemini's Danish responses we performed a correlation analysis between the total word count of each AI generated response and the number of English-only words it contained. This was done for both Gemini 1.5 and Gemini 2.0 across all six prompt categories.

The results showed that gemini 2.0 exhibited significantly stronger correlations compared to Gemini 1.5. In particular, STEM, FAQ, and NON categories in Gemini 2.0 showed statistically significant positive correlations, with Pearson $r = 0.3194$, $r = 0.0970$, and $r = 0.2449$ respectively (all $p < 0.05$). In contrast, Gemini 1.5 only had one significant result, in the RIDDLES category, with a weaker correlation ($r = 0.1749$, $p = 0.0008$).

This suggests that while gemini 1.5 maintains relatively stable output in Danish regardless of prompt length Gemini 2.0 tends to introduce more English words in longer and more complex responses. The difference may reflect changes in model training or English dominant data exposure in newer versions.

Figure 2 and Figure 3 illustrate this contrast between versions. Each blue dot in the scatter plots represents a single AI-generated response. The x-axis shows the total number of words in the response, while the y-axis indicates how many of those words were English. A visible upward trend in Gemini 2.0 (Figure 3) reflects a stronger positive correlation compared to Gemini 1.5 (Figure 2), where no clear pattern is present.

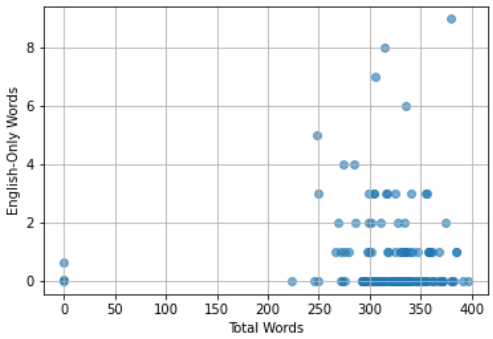

Figure 2: Gemini 1.5 Flash - STEM Category - Answer length vs English usage

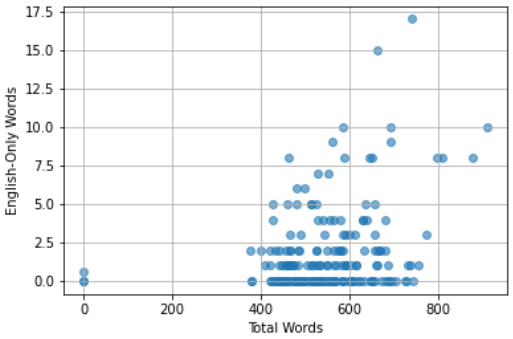

Figure 3: Gemini 2.0 - STEM Category - Answer length vs English usage

The STEM category shows a strong increase in correlation in Gemini 2.0 Flash ($r = 0.3194$), supported by a highly significant p-value.

These findings reinforce the idea that as LLMs become increasingly complex, its ability to differentiate between languages in the output deteriorates, or its decision to keep the languages separate changes. Which

may lead to greater English interference in longer Danish responses, a potential concern in low-resource language contexts.

6.3 Association Rule Mining (ARM)

The ARM analysis revealed a distinct behavioral difference between the two models, directly supporting our primary hypothesis. That at least when it comes to the outputs of the LLM, there is a decline in the use of danish words, instead of English words.

For Gemini 1.5 Flash, the initial analysis generated numerous rules. However, as shown in the processing logs, every single rule was subsequently filtered out for containing only short, high-frequency Danish words (e.g., $\{\texttt{P\_en, P\_i}\} \implies \{\texttt{HAS\_English\_WORD\_IN\_ANSWER}\}$). After applying the word-length filter to focus on more substantive terms, no significant association rules remained. This suggests that for Gemini 1.5, the appearance of English is not strongly tied to any specific, meaningful Danish trigger words in the prompts.

Conversely, the analysis of Gemini 2.0 Flash yielded a significant finding. After the same filtering process, only one rule stood out:

$$\{\text{P\_hvordan}\} \implies \{\text{HAS\_English\_WORD\_IN\_ANSWER}\}$$

- Support: 0.147
- Confidence: 0.570
- Lift: 1.048

This rule indicates that when the Danish word "hvordan" ("how") is present in a prompt, there is a 57% probability that the Gemini 2.0 Flash model will include an English word in its response. Although the Lift is modest (1.048), it shows a positive correlation.

6.4 Categorization of english words

We split up the english words in 5 categories, to better understand how the english words appear in the average response.

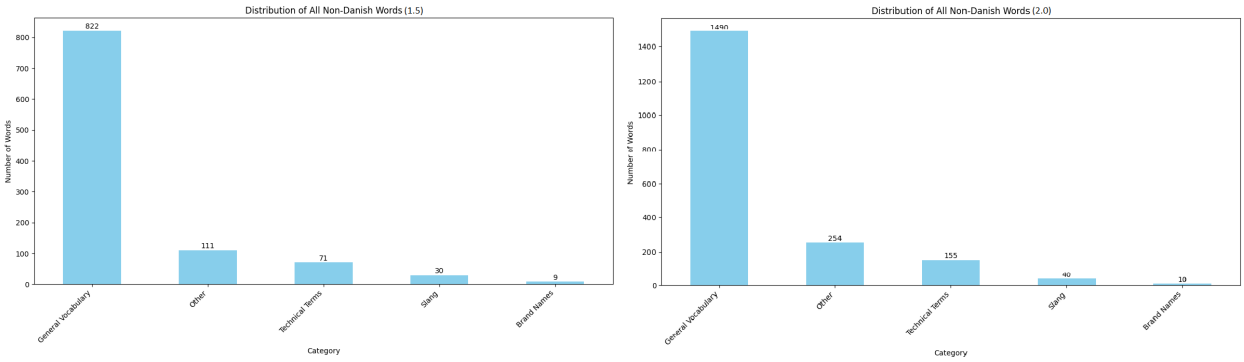

Figure 4: Categories of English words of Google Gemini 1.5 vs. 2.0

Our analysis shows Gemini 2.0 uses more English words than Gemini 1.5, both in absolute terms and proportionally. This corresponds with our findings in regards to correlation between answer length and

English word usage. Calculating the English word rate reveals this increase: Gemini 1.5 averages 0.69 English words per response (1043 words/1502 entries), while Gemini 2.0 averages 0.85 (1949 words/2286 entries).

The categorization of these words, as presented in Table 5, highlights significant behavioral differences between the models.

The categorical distribution reveals several key insights:

- **Technical Language**: Technical Terms saw the most growth, more than doubling in count and increasing proportionally. This supports the idea that Gemini 2.0 leans more on English for complex topics, likely due to its richer English training data.
- **General Vocabulary**: General Vocabulary remains the largest category but slightly declined proportionally, suggesting that growth in English usage is coming more from specialized terms.
- **Other**: The Other category grew notably, pointing to more unpredictable English intrusions, likely encompassing proper nouns, neologisms, and edge cases where English bias surfaces outside defined domains.
- **Slang and Brand Names**: Slang and Brand Names remain minor and declined in share, showing that the increase in English is driven more by technical than informal language.

The increase in English words in Gemini 2.0 is not uniform. The model exhibits a clear trend toward using more technical and specialized English terms, confirming that as models become more powerful, their English-centric training becomes more apparent when generating nuanced responses. This marks a critical evolution in multilingual model behavior.

6.5 Mann-Whitney U Test and means comparison

The Mann-Whitney U test concluded that RIDDLES, NON-STEM and ACADEMIC answers were significantly different and by looking at the graph and the means in Table 3, we see that it is likely that RIDDLES have improved between the two versions while NON-STEM and ACADEMIC has declined.

7 Discussion and Implications

Our findings reveal that English intrusion in Danish language outputs is not just an occasional artifact but a systematic pattern shaped by both model version and input context, seemingly worsening with increased model complexity. While Gemini 1.5 Flash generally maintains Danish linguistic precision, or has sporadic and unpredictable use of English. Gemini 2.0 Flash shows a noticeable increase in English word usage, especially in longer responses and specific categories such as STEM, FAQ, and NON-STEM (see Table 3 and Table 5). This suggests that newer LLM versions, despite improvements in fluency and coherence, may be more prone to code mixing when handling complex or technical prompts. From a sociolinguistic standpoint, this trend is highly noteworthy: repeated exposure to unprompted English elements, especially in educational or informational settings, may gradually normalize anglicisms in everyday Danish usage [Schneider (2022)]. The association rule mining reinforces this concern. While Gemini 1.5 Flash produced no meaningful rules after filtering, Gemini 2.0 Flash yielded a single rule that passed all thresholds: the presence of the word *hvordan* ("how") in the prompt significantly increased the likelihood of English word intrusion in the response. This is important because *hvordan* is a common functional word typically used in prompts asking for explanations or descriptions of processes, suggesting that even ordinary Danish words in explanatory contexts can act as triggers for code-mixing. Furthermore the majority of borrowed English words fall into the general vocabulary category (76.4%), not technical terms (see Figure 4), which challenges the assumption that code-mixing is limited to specialized domains. However, there was also a proportional increase in technical

terms and the "other" category in Gemini 2.0 Flash, suggesting a wider and less predictable spread of English word intrusion across domains.

The correlation analysis further highlights this contrast: Gemini 2.0 Flash exhibited statistically significant positive correlations between response length and English word frequency in multiple categories, for example $r = 0.3194$ in STEM, while Gemini 1.5 Flash showed only one significant but weaker correlation in RIDDLES category ($r = 0.1749$; see Table 5, Figures 2–3). While these correlation findings offer valuable insights into the differing behaviors of the two models, they are further contextualized by the results of the ANOVA analysis. Although the ANOVA analysis encountered some complications in regards to the non-normally distributed residuals [Shapiro & Wilk (1965)], the implications still somewhat aligned with our other findings.

Taken together, these patterns highlight how LLMs might unintentionally contribute to gradual erosion of linguistic boundaries. As generative AI tools become more integrated into public services, such as chatbots, translation tools, or educational platforms, the risk of reinforcing English dominance in multilingual contexts increases. This has implications not only for Danish but for any low-resource or minority languages .

The conclusion drawn from this specific result is critical: *hvordan* prompts inherently ask for explanations, instructions, or descriptions of processes. The fact that this word is a lone, significant trigger for English inclusion in the newer model strongly suggests that as response complexity increases, Gemini 2.0 Flash is more likely to draw upon English terminology. This provides concrete evidence for the conclusion that newer models, despite their advanced capabilities, exhibit a greater tendency to mix languages in more elaborate or technical Danish-language contexts.

8 Conclusion and Future Work

This study examined the amount of unprompted English words appear in Danish language outputs generated by the two versions of Google's Gemini large language models [Google (2025)]. Through quantitative and linguistic analysis, we found that Gemini 2.0 Flash introduces statistically significant more unprompted English responses than Gemini 1.5 Flash, particularly in longer and more complex prompts.

These findings highlight the risk that generative AI tools, despite being multilingual, may unintentionally reinforce the presence of English in low-resource language settings. This has important implications for linguistic diversity and digital language development.

The analysis also showed that the type of English words changes between versions. Gemini 2.0 Flash includes more technical terms and a broader range of general vocabulary, suggesting a shift toward deeper integration of English across domains. The association rule mining showed that even a common Danish word like "hvordan" can strongly predict English word usage in responses from Gemini 2.0 Flash. This provides concrete evidence that as the model responds to more complex prompts, it tends to draw more heavily on its English training data, resulting in increased codemixing.

Taken together, these patterns raise concern that as LLMs become more advanced, their reliance on English may grow—not just in technical contexts but also in everyday responses, making English borrowings more common in general Danish output.

Finally, future research should aim to understand how often Danish speakers mix English in real life, and to see how this compares to what happens in other languages, such as Norwegian, French etc. This would help contextualize LLM behavior within broader sociolinguistic norms and guide efforts toward more equitable and culturally sensitive multilingual AI systems. Also, in hindsight we should have conducted the Pearson correlation for prompt length rather than response length, which is relevant for future work.

References

Technology Networks. (2023). *Mann-Whitney U Test: Assumptions and Example*. Retrieved June 8, 2025, from https://www.technologynetworks.com/informatics/articles/mann-whitney-u-test-assumptions-and-example-363425

Google. (2025). *Generative AI Developer Guide*. Retrieved June 8, 2025, from https://ai.google.dev

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611.

Gottlieb, J. (2004). *Modersmålet i fare?* Retrieved June 8, 2025, from https://web.archive.org/web/20071006142535/http://www.modersmaalet.dk/gottlieb2004.pdf

Smith, A., & Jones, B. (2025). Evaluating GPT-4 and Gemini 1.5 Pro on multilingual tasks. *arXiv*. Retrieved June 8, 2025, from https://arxiv.org/html/2502.15603v1

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. doi:10.1145/3442188.3445922

Schneider, B. (2022). Multilingualism and AI: The regimentation of language in the age of digital capitalism. *Signs and Society, 10*(3), 362–387. doi:10.1086/721757

Appendix

Github Repo

DISCLOSURE: AI was used as help in debugging the code.
https://github.com/Nunk-sketch/stat-eval-2

Mann-Whitney U Test by Category

Category	p-value
ACADEMIC	0.000492
CREATIVE PROMPTS	0.185114
FAQ	0.438541
MULTIPLE	0.431021
NON	0.000073
RIDDLES	0.019560
STEM	0.065464

Table 7: Mann-Whitney U Test p-values by Category

Distribution of words sorted in categories for Gemini 1.5 Flash

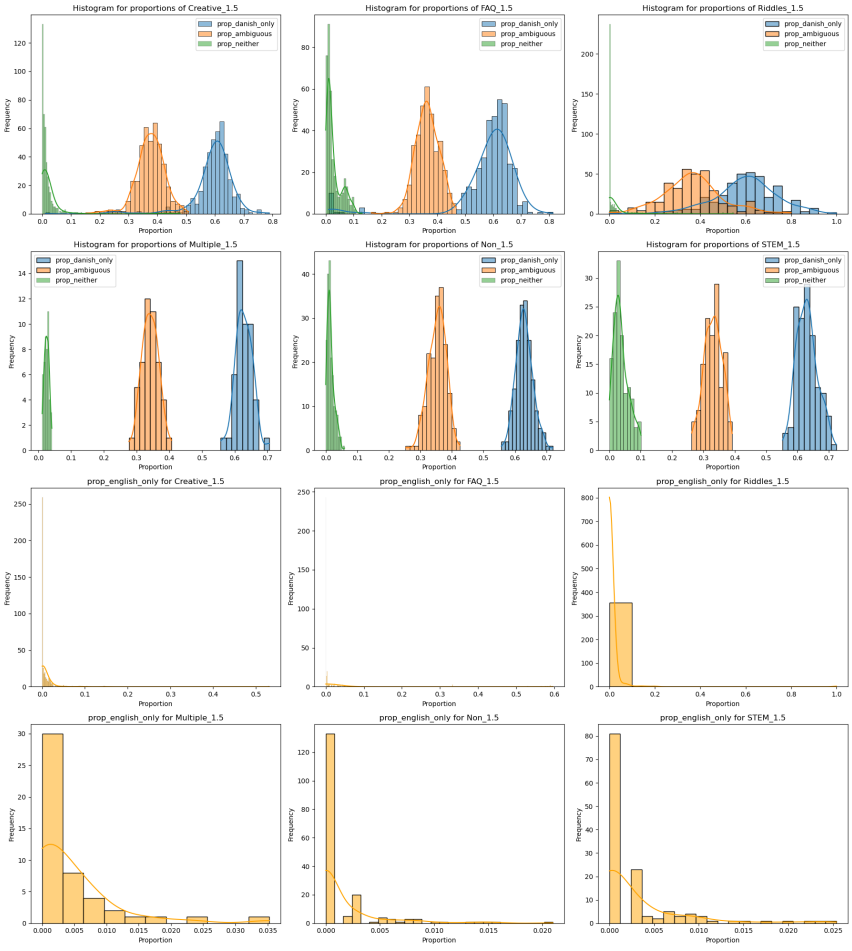

Figure 5: all histograms 1.5 proportion

Distribution of words sorted in categories for Gemini 2.0 Flash

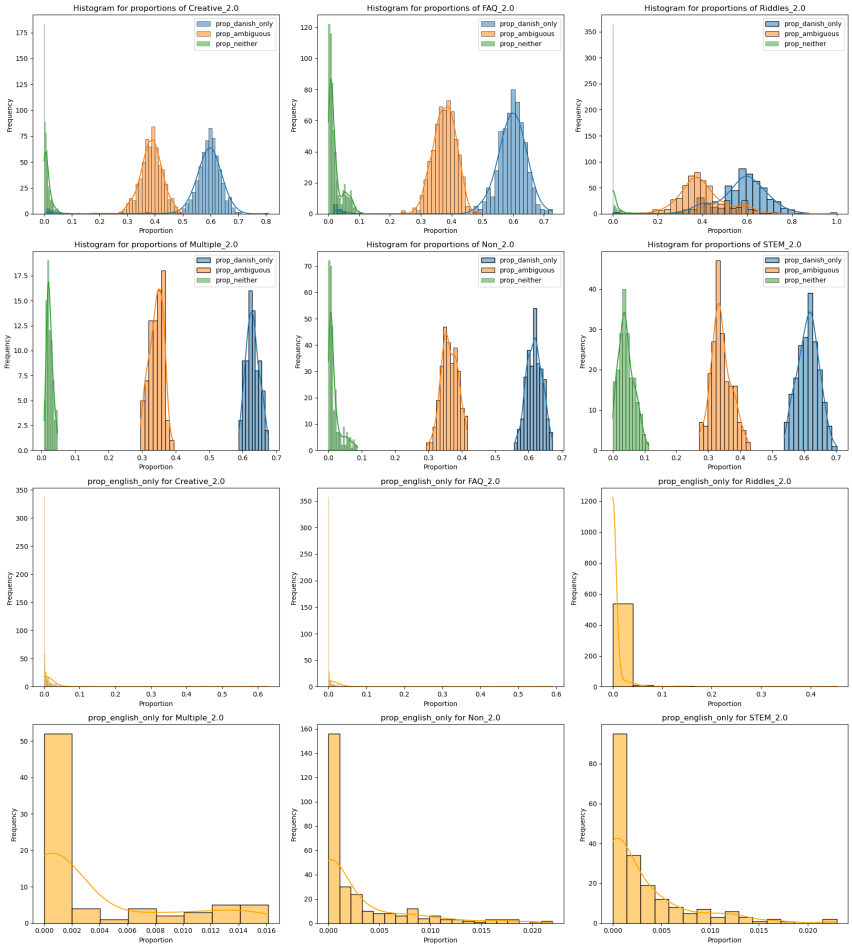

Figure 6: all histograms 2.0 proportion