PROJECT STATISTICS: AI EVALUATION

MATTIS KRAGH (S246222), RO NANAK PRASAD LACOUL (S246152), ANTONIJS BOLSAKOVS (S225124),
CHRISTOPHER ALEXANDER HOLM KJÆR (S234744)

June 10, 2025

1 Motivation and goal

LLMs are increasingly used across languages, but English remains dominant in both training and output [Bender et al., 2021]. This poses challenges for smaller languages like Danish, where AI-generated content often includes unprompted English terms [Gottlieb, 2004].

This project investigates whether such interference is systematic, quantifiable, and culturally significant. We aim to assess the extent and patterns of English in Danish AI outputs and reflect on implications for linguistic diversity.

2 Problem statement

As large language models (LLMs) like Google Gemini 2.5 become central to digital communication, information access, and cultural exchange, questions arise about the dominance of English in their outputs [Schneider, 2022]. Although these models are trained on multilingual data, English remains the primary language, which may shape the way other languages are represented and used [Bender et al., 2021].

This project explores how Google Gemini 1.5/2.0 Flash handles Danish—a comparatively low-resource language. When Danish users interact with these models, they often receive responses that include English words or phrases, even when the input is written entirely in Danish. This suggests a broader pattern of English influence in digital communication across non-English languages [Schneider, 2022].

Our aim is to investigate the frequency and form of English intrusions in Danish-language outputs. By generating a range of Danish prompts and analyzing the model's responses, we seek to measure the extent of this interference and reflect on its potential impact. At the core of this inquiry is a cultural and linguistic concern: Does repeated exposure to English in AI outputs contribute to the erosion of Danish, normalizing anglicisms and subtly reshaping language use among Danish speakers [Gottlieb, 2004]?

2.1 Sub questions

- What proportion of non-English prompts contain English elements?
- Which categories of prompts (e.g., technical instructions, casual conversation, creative tasks, or academic questions) show the highest average frequency of English content in Danish-language input?
- Is there statistically significant correlation between prompt length (either in tokens or words) and the occurrence of english words?
- Which specific Danish words or expressions in prompts are statistically associated with a higher likelihood of English word-borrowing, as identified through association rule mining techniques?
- What types of English words are used in danish language prompts, and how can they be categorized (e.g., as technical terms, slang, general vocabulary, or brand names)?

3 Potential methodology of the evaluation pipeline

The evaluation process consists of the following steps:

- **Prompt Design:** We will create a diverse set of prompts in Danish, divided into thematic categories (e.g. creative prompts, riddles, academic questions).

- **Response Collection:** For each prompt, we will query Gemini 1.5 and Gemini 2.0 Flash separately and record the model's output.
- **Annotation and Detection:** Using both automated text processing (e.g., regular expressions, language detection libraries) and (limited amount of) manual verification, we will identify and tag English content embedded in Danish responses.
- **Quantitative Analysis:** We will compute the proportion of English words, the average frequency of interference across prompt types, and perform correlation analysis (e.g., Pearson or Spearman) to test dependencies on prompt length or category.
- **Statistical Testing:** Where appropriate, we will apply significance testing (e.g., chi-square tests or t-tests) to evaluate differences between the two sets of data or results.
- **Linguistic Categorization:** We will attempt to classify observed English insertions according to type (e.g., nouns vs. verbs, slang vs. formal terms), which will help contextualize their presence.

4 Exptected outcome

We expect to find that English appears in a measurable portion of outputs generated from Danish prompts, with a higher rate in Gemini 1.5 than in Gemini 2.0 due to improved multilingual training [Smith & Jones, 2025]. We also anticipate that certain categories of prompts (e.g., technical or digital topics) will exhibit higher levels of English interference, reflecting domain-specific vocabulary. Additionally, the results may reveal whether AI-generated anglicisms are isolated or systematic, and whether certain prompt structures consistently trigger them. The broader outcome is a deeper understanding of how multilingual AI systems reinforce or reduce language dominance in practice [Schneider, 2022].

5 Outline

1. Introduction
2. Problem and Motivation
3. Methodology
4. Data Collection
5. Analysis
6. Results
7. Discussion and Implications
8. Conclusion and Future Work

6 Discussion points

One important point of discussion is whether the increasing presence of English in prompts written in other languages represents a threat to linguistic diversity [Gottlieb, 2004]. As English continues to dominate in digital spaces, including AI interactions, speakers of smaller or non-global languages may gradually adopt more English vocabulary, even when communicating in their native language. This phenomenon raises ethical and cultural concerns regarding the responsibility of AI developers to support linguistic plurality and resist reinforcing existing language hierarchies [Bender et al., 2021, Schneider, 2022].

References

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