LLMs vs. Traditional Sentiment Tools in Psychology: An Evaluation on Belgian-Dutch Narratives

Ratna Kandala, Katie Hoemann
Department of Psychology
University of Kansas
Larewnce, KS 66045
n038k926@ku.edu, hoemann@ku.edu

Abstract

Understanding emotional nuances in everyday language is crucial for computational linguistics and emotion research. While traditional lexicon-based tools like LIWC and Pattern have served as foundational instruments, Large Language Models (LLMs) promise enhanced context understanding. We evaluated three Dutch-specific LLMs (ChocoLlama-8B-Instruct, Reynaerde-7B-chat, and GEITje-7B-ultra) against LIWC (27) and Pattern (9) for valence prediction in Flemish, a low-resource language variant. Our dataset comprised approximately 25,000 spontaneous textual responses from 102 Dutch-speaking participants, each providing narratives about their current experiences with self-assessed valence ratings (–50 to +50). Surprisingly, despite architectural advancements, the Dutch-tuned LLMs underperformed compared to traditional methods, with Pattern showing superior performance. These findings challenge assumptions about LLM superiority in sentiment analysis tasks and highlight the complexity of capturing emotional valence in spontaneous, real-world narratives. Our results underscore the need for developing culturally and linguistically tailored evaluation frameworks for low-resource language variants, while questioning whether current LLM fine-tuning approaches adequately address the nuanced emotional expressions found in everyday language use.

1 Introduction

The increasing use of Large Language Models (LLMs) in everyday applications has raised important questions about their ability to understand emotional nuances, particularly as users seek empathetic interactions with AI systems. Current LLMs lack psychological mechanisms for genuine empathy (35), making culturally and linguistically tailored valence detection critical, especially for low-resource languages like Flemish (Belgian-Dutch). Traditional lexicon-based tools like LIWC (29) and Pattern (9) have dominated valence analysis through predefined word-emotion mappings. While LIWC excels with validated dictionary categories and Pattern offers robust rule-based analysis, both struggle with spontaneous, context-rich narratives. For example, LIWC would misclassify sarcastic expressions like "This *amazing* Monday has me stuck in traffic for hours - what a *joy*!" as positive, failing to detect contextual frustration. Most sentiment analysis research focuses on high-resource languages like English (16), while Flemish remains underrepresented despite being spoken by 6.5 million Belgians (20). Flemish exhibits distinct features from standard Dutch, including regional vocabulary ("fuif" for "party"), divergent pronouns ("gij" vs. "jij"), and multilingual influences, challenging existing tools. This study addresses three critical gaps: linguistic inequity in NLP research, ecological validity (moving beyond constrained social media data (33; 34)), and methodological rigor using self-reported rather than annotator-based ground truth. We compare lexicon-based tools (LIWC, Pattern) with Dutch-tuned LLMs (ChocoLlama-8B-Instruct, Reynaerde-7B-chat, GEITje-7B-ultra) using nearly 25,000 open-ended Flemish narratives from 102 participants collected via

ambulatory assessment. Participants provided real-time experiences with self-rated valence scores (–50 to +50). Our findings reveal that state-of-the-art Dutch LLMs underperform lexicon-based methods, underscoring needs for linguistically tailored models and equitable AI development that respects linguistic diversity in applications from mental health interventions to social science research.

2 Prior Work

Sentiment analysis, detecting emotional valence in language, is widely applied across healthcare, customer service, and misinformation detection, with traditional lexicon-based tools like LIWC (29) and Pattern (9) dominating psychological research due to their interpretability, though they lack contextual awareness and struggle with sarcasm and domain-specific cues (22). While recent studies highlight LLMs' potential for English valence detection (19), their performance in low-resource languages remains understudied, with untuned models often failing to capture emotion intensity due to English-centric training data (18; 16). For Flemish, spoken by 6.5 million Belgians with distinct linguistic features, sentiment analysis research is sparse, with existing work like Reusens (33) limited to social media data that distorts emotional expression through character limits and platform constraints (34). We address these gaps by evaluating three Dutch-tuned LLMs against LIWC and Pattern on nearly 25,000 open-ended Flemish narratives with self-reported valence scores, using ambulatory assessment data that preserves linguistic authenticity and provides ecologically valid ground truth compared to external annotator approaches (26; 12).

3 Lexicon-Based Text Analyses and LLMs Chosen

Lexicon-based methods employ dictionaries to detect sentiments, valued for their interpretability and ease of implementation. Linguistic Inquiry and Word Count (LIWC) (27), adapted for Dutch (5), maps words across 90+ categories including emotional tone (PosEmo and NegEmo). The Dutch version contains 6,614 words (1,227 PosEmo, 1,474 NegEmo), estimating valence by calculating percentage of dictionary-matched words. However, as a proprietary tool, LIWC lacks transparency and doesn't incorporate negation or sarcasm detection rules. Pattern.nl (9) is an open-source Python package providing "plug-and-play" sentiment analysis without machine learning expertise for psychology researchers. It uses a handcrafted lexicon of 3,304 Dutch lemmas (97% adjectives) with polarity values [-1,1]. The algorithm performs chunking, chunk-level scoring with intensifier/negation rules, and sentence-level aggregation. However, the lexicon, compiled from product reviews, may score affective texts as neutral when words fall outside this domain. We selected three open-weight Dutch-tuned LLMs for valence estimation: ChocoLlama-8B-Instruct, GEITje-7B-Ultra, and Reynaerde-7B-Chat, chosen for architectural diversity and data privacy considerations. ChocoLlama-8B-Instruct (20) is built on Llama-3-8B with rank-16 LoRA on 32B Dutch tokens, followed by supervised fine-tuning and DPO alignment. GEITje-7B-Ultra (38) is based on Mistral 7B with full-parameter pretraining on 10B Dutch tokens, followed by SFT and DPO alignment, reported as the best GEITje variant. Reynaerde-7B-Chat (32) uses rank-8 QLoRA adapters on Llama-2-7B-hf, fine-tuned on 400K Dutch chat pairs with DPO alignment for safe dialogue. Both GEITje and Reynaerde have outperformed earlier ChocoLlama variants on Dutch benchmarks, though their psychological research efficacy remains unvalidated (20).

4 Methodology

4.1 Data

The dataset comprises 24,854 open-ended Dutch-language text entries (medium-scale) collected longitudinally over 70 days from 102 participants (age : 18–65, μ = 26.47, σ = 8.87) in Belgium. Participants were native Dutch speakers with smartphone access, recruited via online/posted flyers and referrals (most were native Belgians, a few were from the Netherlands and living in Belgium). Using a dedicated app, they received four daily prompts asking (in Dutch), "What is going on now or since the last prompt, and how do you feel about it?" Responses were either short text snippets (3–4 sentences) or 1-minute voice recordings, yielding a temporally structured dataset with potential for time-series or multimodal analysis (text + audio). For robustness, eligibility criteria ensured linguistic consistency (native speakers) and device reliability (smartphone ownership). Further methodological

details, including ethical protocols, are provided in the Supplementary Material. The dataset can be made available to interested persons after acceptance (in compliance with GDPR). Nevertheless, the users' self-reported valence scores, the ratings by each of these tools/models will be made available online.

4.2 Prompt for the LLMs

To enable sentiment analysis on this dataset, we designed a standardized prompt to elicit valence ratings from various Dutch-language large language models (LLMs). We experimented with both Dutch and English prompts but did not observe any significant improvement in LLM performance with the Dutch prompts. Similarly, we found no notable difference between zero-shot and few-shot settings (see Supplementary Materials for details). So we proceeded ahead with the English prompt:

"You are a Dutch language expert analyzing the valence of Belgian Dutch texts. Participants responded to: 'What is going on now or since the last prompt, and how do you feel about it?' Carefully read the response of the participant: {text}. Your task is to rate its sentiment from 1 (very negative) to 7 (very positive). Return ONLY a single numerical rating enclosed in brackets, e.g. [X], with no additional text.
Output Format: [number]"

The placeholders {text} are filled with the texts.

5 Results: Performance of LLMs versus Lexical Tools

This section presents the results of the models and their comparison with self-reported valences of the users. Coverage analysis revealed stark differences between methods, with lexicon-based tools (LIWC and Pattern) achieving near-complete coverage at 99.9% (24,848/24,852 texts), while Dutch-tuned LLMs showed substantial variability: ChocoLlama-8B-Instruct processed 69.9% of texts, GEITje-7B-ultra only 38.0%, and Reynaerde-7B-chat a mere 1.8%. These coverage gaps in LLMs risk introducing selection bias and complicate fair performance evaluations compared to the universal processing capability of lexicon tools. To assess alignment with human judgment, we evaluated how well each model's predictions correlated with users' self-reported valence ratings (ranging from –50 to +50) using both Pearson's r for linear association and polyserial correlation for ordinal-continuous relationships. Higher correlations indicated that model predictions more closely tracked the writers' own emotional assessments of their narratives.

Pattern demonstrated superior lexicon-based performance with correlations of r = 0.31 (both Pearson and Polyserial), outperforming LIWC15-PosEmo (r = 0.21/0.23) and LIWC15-NegEmo (r = -0.23/-0.23) across nearly all 24,852 texts. Among LLMs, ChocoLlama-8B-Instruct showed the strongest performance on its 17,378 processed texts (Pearson r = 0.35, Polyserial r = 0.40), significantly outperforming Pattern ($t = 4.02$, $p < 0.001$) on this subset. GEITje-7B-ultra achieved similar correlations (r = 0.35/0.44) across 9,445 texts but did not significantly differ from Pattern ($t = -1.20$, $p = 0.23$), while Reynaerde-7B-chat underperformed with correlations of only r = 0.18/0.24 on 446 texts, showing no significant advantage over lexicon methods. The key trade-off emerged between LLM accuracy and coverage: ChocoLlama and GEITje achieved higher correlations than lexicon tools on their processed subsets, but their limited coverage (69.9% and 38.0% respectively) versus Pattern's near-universal coverage (99.9%) raises questions about practical applicability and potential selection bias.

Table 1: Correlation and coverage metrics across sentiment models

Model	Coverage (N / %)	Pearson r	Polyserial r
LIWC (posemo)	24,848 / 99.9%	0.21	0.23
LIWC (negemo)	24,848 / 99.9%	-0.23	-0.23
Pattern.nl	24,848 / 99.9%	0.31	0.31
ChocoLlama-8B-Instruct	17,378 / 69.9%	0.35	0.40
GEITje-7B-ultra	9,445 / 38.0%	0.35	0.44
Reynaerde-7B-chat	446 / 1.8%	0.18	0.24

6 Discussion

Our evaluation of Dutch-finetuned LLMs against lexicon-based tools for Flemish valence detection revealed that while models like ChocoLlama-8B-Instruct and GEITje-7B-ultra showed moderate correlations with human ratings, they suffered from incomplete coverage and systematic domain mismatch between training data and real-world narratives. These findings challenge assumptions about LLM superiority in low-resource settings. The underperformance stems from fundamental training data limitations. Major LLMs exhibit severe English-centric bias—Llama-2's pretraining comprises 89.7% English content, while GPT-3 contains roughly 92.65% English tokens (15). Although adaptation efforts like ChocoLlama incorporated 32 billion Dutch tokens for continued pretraining (20), overreliance on formal corpora (legal documents, social media) creates models ill-suited for colloquial, first-person narratives characteristic of daily emotional expression. This domain mismatch undermines valence accuracy because LLMs lack exposure to lexical and pragmatic cues signaling emotion in everyday language (4). Conversely, LIWC and Pattern.nl, grounded in psycholinguistic principles and manually curated word lists, retain sensitivity to informal affective expressions (28; 13). Pattern.nl's superior performance likely reflects its composite polarity scoring versus LIWC's raw emotion-word counts. Addressing LLM limitations requires augmenting training data with narrative-style corpora, creating hybrid lexicon-LLM models, and developing standardized Flemish valence benchmarks (3). Until LLMs overcome domain mismatches, lexicon tools remain indispensable for scalable, ecologically valid valence analysis in low-resource languages like Flemish.

7 Conclusion and Limitations

This study demonstrates that lexicon-based tools like Pattern.nl remain the gold standard for valence analysis in low-resource Flemish, outperforming Dutch-tuned LLMs in coverage and reliability despite contextual limitations. Our findings highlight critical gaps in current LLM approaches: incomplete coverage, domain mismatch between formal training data and colloquial narratives, and systematic underrepresentation of Flemish linguistic features. Several limitations constrain our findings. First, Dutch and Flemish remain critically under-represented in training data and available LLMs, reflecting broader systemic gaps in low-resource language NLP (20). Our evaluation of three Dutch-tuned models does not account for larger architectures (e.g., LLaMA-2 70B) or hybrid approaches that may outperform current benchmarks. Second, while our corpus of nearly 25,000 narratives provides ecological validity, reliance on a single prompt structure risks conflating valence expression with elicitation method, limiting generalizability. Finally, lexicon tools, though reliable, lack contextual nuance for resolving sarcasm or cultural idioms and cannot replicate human empathy (25). Moving forward, establishing standardized Flemish valence benchmarks grounded in manually annotated narratives will enable targeted improvements in data augmentation, bias mitigation, and hybrid modeling. The integration of LLMs with lexicon-driven frameworks may bridge the gap between scalable automation and culturally grounded emotion research. Until such synergies are realized, lexicon tools will continue to anchor valence analysis in low-resource contexts, ensuring both methodological rigor and practical applicability for psychological research and mental health applications.

References

[1] Alexandridis, G., Varlamis, I., Korovesis, K., Caridakis, G., and Tsantilas, P. (2021). A survey on sentiment analysis and opinion mining in Greek social media. *Information*, 12:331.

[2] Ali, H., Hashmi, E., Yildirim, S. Y., and Shaikh, S. (2024). Analyzing Amazon products sentiment: a comparative study of machine and deep learning, and transformer-based techniques. *Electronics*, 13(7):1305.

[3] Augustyniak, Ł., Woźniak, S., Gruza, M., Gramacki, P., Rajda, K., Morzy, M., and Kajdanowicz, T. (2023). Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark. *arXiv preprint arXiv:2306.07902*.

[4] Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, Bollywood, boom-boxes: domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.

[5] Boot, P., Zijlstra, H., and Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, 6(1):65–76.

[6] Bordia, S. and Bowman, S. R. (2019). Identifying and Reducing Gender Bias in Word-Level Language Models. In Kar, S., Nadeem, F., Burdick, L., Durrett, G., and Han, N.-R. (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, MN. Association for Computational Linguistics. https://aclanthology.org/N19-3002/. 10.18653/v1/N19-3002

[7] Chun, J. (2021). SentimentArcs: a novel method for self-supervised sentiment analysis of time series shows SOTA transformers can struggle finding narrative arcs. *arXiv preprint arXiv:2110.09454*.

[8] Dashtipour, K., Gogate, M., Adeel, A., Larijani, H., and Hussain, A. (2021). Sentiment analysis of Persian movie reviews using deep learning. *Entropy*, 23:596.

[9] De Smedt, T. and Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, 13(1):2063–2067.

[10] Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., Mitchell, N. J., Ong, D. C., Dweck, C. S., Gross, J. J., and Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

[11] Fiok, K., Karwowski, W., Gutierrez, E., and Wilamowski, M. (2021). Analysis of sentiment in tweets addressed to a single domain specific Twitter account: comparison of model performance and explainability of predictions. *Expert Systems with Applications*, 186:115771.

[12] Frisina, P. G., Borod, J. C., and Lepore, S. J. (2004). A meta-analysis of the effects of written emotional disclosure on the health outcomes of clinical populations. *The Journal of Nervous and Mental Disease*, 192(9):629–634.

[13] Lorenzo Gatti and Judith van Stegeren. Improving Dutch sentiment analysis in Pattern. *Computational Linguistics in the Netherlands Journal*, 10:73–89, December 2020. (Submitted October 2020).

[14] Hutto, C. J. and Gilbert, E. (2014). VADER: a parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

[15] Li, Z., Shi, Y., Liu, Z., Yang, F., Payani, A., Liu, N., and Du, M. (2024). Quantifying multilingual performance of large language models across languages. *arXiv preprint arXiv:2404.11553*.

[16] Ligthart, A., Catal, C., and Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, 54:4997–5053.

[17] Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., and Ananiadou, S. (2023). Emotion detection for misinformation: a review. *arXiv preprint arXiv:2311.00671*.

[18] Liu, Z., Yang, K., Xie, Q., Zhang, T., and Ananiadou, S. (2024). EmoLLMs: a series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

[19] Martínez, G., Molero, J. D., González, S., Conde, J., Brysbaert, M., and Reviriego, P. (2025). Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, 57:5. https://doi.org/10.3758/s13428-024-02515-z

[20] Meeus, M., Rathé, A., Rémy, F., Delobelle, P., Decorte, J.-J., and Demeester, T. (2024). ChocoLlama: lessons learned from teaching llamas Dutch. *arXiv preprint arXiv:2412.07633*.

[21] Mishra, K., Priya, P., Burja, M., and Ekbal, A. (2023). e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13952–13967.

[22] Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

[23] Muhammad, S., Abdulmumin, I., Ayele, A., Ousidhoum, N., Adelani, D., Yimam, S., Ahmad, I., Beloucif, M., Mohammad, S., Ruder, S., Hourrane, O., Jorge, A., Brazdil, P., Ali, F., David, D., Osei, S., Shehu-Bello, B., Lawan, F., Gwadabe, T., Rutunda, S., Belay, T., Messelle, W., Balcha, H., Chala, S., Gebremichael, H., and Opoku, B. (2023). AfriSenti: a Twitter sentiment analysis benchmark for African languages. In *Proceedings of EMNLP 2023*.

[24] Obradovich, N., Khalsa, S. S., Khan, W. U., et al. (2024). Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry Neuroscience*, 2:8.

[25] Pataranutaporn, P., Liu, R., Finn, E., et al. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5:1076–1086.

[26] Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, 8(3):162–166.

[27] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum Associates, Mahwah, NJ.

[28] Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2007). *Linguistic Inquiry and Word Count: LIWC2007: operator's manual*. Austin, TX: LIWC.net.

[29] Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates. `https://www.liwc.net`

[30] Qiu, J., Lam, K., Li, G., et al. (2024). LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6:1418–1420.

[31] Radaideh, M. I., Hwang, O. H., and Radaideh, M. I. (2024). Fairness and social bias quantification in large language models for sentiment analysis. SSRN.

[32] ReBatch. (2024). Reynaerde-7B-Chat. *Hugging Face*.

[33] Reusens, M., Reusens, M., Callens, M., Vanden Broucke, S., and Baesens, B. (2022). Benchmark study for Flemish Twitter sentiment analysis. SSRN. `https://ssrn.com/abstract=4096559`

[34] Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082. `https://doi.org/10.1111/ajps.12103`

[35] Shteynberg, G., Halpern, J., Sadovnik, A., et al. (2024). Does it matter if empathic AI has no empathy? *Nature Machine Intelligence*, 6:496–497.

[36] Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

[37] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., *et al.* (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*. `https://arxiv.org/abs/2307.09288`

[38] Vanroy, B. (2024). GEITje 7B Ultra: a conversational model for Dutch. *arXiv preprint arXiv:2412.04092*.

[39] Varghese, C., Harrison, E. M., O'Grady, G., et al. (2024). Artificial intelligence in surgery. *Nature Medicine*, 30:1257–1268.

[40] Zhang, W., Deng, Y., Liu, B., Pan, S., and Bing, L. (2024). Sentiment analysis in the era of large language models: a reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

A Supplementary Material

A.1 Data Preparation

Participants were recruited from the community via flyers, online advertisements, and word-of-mouth referrals. Eligibility criteria included being native Dutch speakers, residing in Belgium, being at least 18 years of age, and owning a fully functional smartphone. Interested individuals completed an initial eligibility survey, with eligibility criteria verified during a subsequent online introduction session.

Participants received compensation of up to €250 for participating in a 70-day (10-week) experience sampling protocol, accompanied by biweekly online surveys. The breakdown included €0.50 per completed experience sampling prompt (maximum of 280 prompts or €140 total), €10 for each short survey at 2, 4, 6, and 8 weeks (maximum €40), and €15 for each long survey at baseline and 10 weeks (maximum €30). An additional bonus of €40 was given to participants completing at least 60 days. Payment was issued in full upon study completion. Continued participation required completion of at least 75% of prompts, with verbal responses having a minimum length of 25 words. Regular compliance checks and biweekly summary reports were provided to participants, with reminders issued as necessary.

Initially, 115 participants enrolled (age range: 18–65, M = 27.26, SD = 9.86; gender: 58 female, 56 male, 1 other). Of these, 10 voluntarily withdrew, and 3 were dismissed due to poor compliance (response rates below 50%). Thus, 102 participants completed the study (age range: 18–65, M = 26.47, SD = 8.87; 52 women, 49 men, 1 other).

Ethical approval was obtained from the KU Leuven Social and Societal Ethics Committee (SMEC), protocol G-2023-6379-R3(AMD). Data collection spanned from August 2023 to July 2024. All study materials and instructions were administered in Dutch.

A.1.1 Procedure and Materials

Participants engaged in a 70-day experience sampling protocol via a dedicated mobile application (m-Path; Mestdagh et al., 2023), receiving four prompts daily at pseudorandom intervals between 9 AM and 9 PM, spaced at least one hour apart. At each prompt, participants responded to "What is going on now or since the last prompt, and how do you feel about it?" Responses were typically recorded as 1-minute voice messages but could optionally be typed (3-4 sentences). Participants then rated their current emotional valence on a slider scale ranging from -50 (very unpleasant) to +50 (very pleasant). Simultaneously, m-Path recorded sensor data (GPS coordinates, ambient noise, step counts, and app usage for Android devices).

All assessment tools were validated previously in Dutch-speaking samples. Participants completed additional online surveys after 2, 4, 6, and 8 weeks, and a comprehensive survey at 10 weeks, which included validation questions about recent visited locations.

Participants received biweekly compliance summary reports via email, detailing prompt completion rates, description length adequacy, modality (voice vs. text), valence ratings, linguistic content trends (using LIWC 2015 Dutch translation), and accumulated compensation. These reports also indicated whether participants earned bonuses and their overall compensation.

Voice-recorded responses were automatically transcribed using a proprietary algorithm developed by the Department of Electrical Engineering (ESAT) at KU Leuven (Tamm et al., 2024). Transcripts and typed responses were integrated into a unified data file, aligned manually with online survey data into five study intervals: days 1–14, 15–28, 29–42, 43–56, and 57–70. Missing survey data were noted where applicable.

A.1.2 Analytic Tools and Experimental Setup

Textual responses were analyzed using Linguistic Inquiry and Word Count (LIWC) with its Dutch lexicon, generating category-specific linguistic scores for each response. Additionally, the Pattern.nl sentiment analysis algorithm provided continuous valence estimates from -1 (very negative) to +1 (very positive), accounting for word context and grammatical function. Model evaluations were performed using a NVIDIA RTX-5000 GPU and implemented in PyTorch, with models obtained via Hugging Face APIs.

A.2 Single Participant (Pilot Study)

We conducted a preliminary pilot study involving one participant's data from our dataset to identify the best performing models. We also compared the performances of these models with English prompt and Dutch prompt.

English Prompt: "You are a Dutch language expert analyzing the valence of Belgian Dutch texts. Participants responded to: "What is going on now or since the last prompt, and how do you feel about it?" Carefully read the response of the participant: "{text}".
Your task is to rate its sentiment from 1 (very negative) to 7 (very positive).
Return ONLY a single numerical rating enclosed in brackets, (e.g. [X]), with no additional text, explanations, or formatting.
Output Format: [number] . Replace "number" with the integer score (1-7)."

Dutch Prompt: "Je bent een Nederlandse taalexpert die de valentie van Belgisch Nederlandse teksten analyseert. Deelnemers reageerden op:
"Wat speelt er nu of sinds de vorige beep en hoe voel je je daarover?". Lees zorgvuldig het antwoord van de deelnemer: "{text}".
Het is jouw taak om het sentiment te beoordelen van 1 (zeer slecht) tot 7 (zeer goed). Geef ALLEEN een cijfer tussen haakjes (bijv. [X]), zonder extra tekst, uitleg of opmaak. Niet uitleggen. Uitvoerformaat: [getal]. Vervang "getal" door de gehele score (1-7)."

As illustrated in Table 1, we selected ChocoLlama-8B-Instruct, GEITje-7B-ultra, Reynaerde-7B-Chat, and Llama 2-7B. We also ran the dataset with LIWC15 and Pattern. The models returned values for all the texts. Among these LLMs, when prompted in English, ChocoLlama-8B-Instruct achieved the highest polyserial correlation (r = 0.55) with the user's ratings followed by GEITje-7B-ultra (r = 0.42) and Reynaerde-7B-Chat (r = 0.33). Surprisingly, Llama-2-7B returned no values for all the texts when prompted in English but performed better than GEITje-7B-ultra (r = 0.0002) and Reynaerde-7B-Chat (r = 0.42). To maintain uniformity, we decided to proceed with only English prompt for further experiments. Between LIWC15 and Pattern.nl, LIWC15 showed strong negative (negemo: r = -0.54) and positive (posemo: r = 0.41) correlations, and Pattern.nl showed positive correlation with the user's ratings (r = 0.44).

Table 2: Correlation coefficients across models and prompts

Model (variable)	**Prompt**	**Pearson** r	**Polyserial** r
LIWC15 (posemo)	–	0.41	–
LIWC15 (negemo)	–	-0.54	–
Pattern.nl	–	0.44	–
ChocoLlama-8B-Instruct	English	0.47	0.55
ChocoLlama-8B-Instruct	Dutch	0.18	0.27
Llama2-7B	English	N/A	N/A
Llama2-7B	Dutch	0.37	0.46
GEITje-7B-ultra	English	0.33	0.42
GEITje-7B-ultra	Dutch	0.0001	0.0002
Reynaerde-7B-Chat	English	-0.1	0.33
Reynaerde-7B-Chat	Dutch	0.2	0.42

In the following sections, we present the additional experiments conducted and the results observed for the entire dataset using the English prompt:

A.3 Distribution of Ratings (All Participants/Users, LIWC, Pattern.nl)

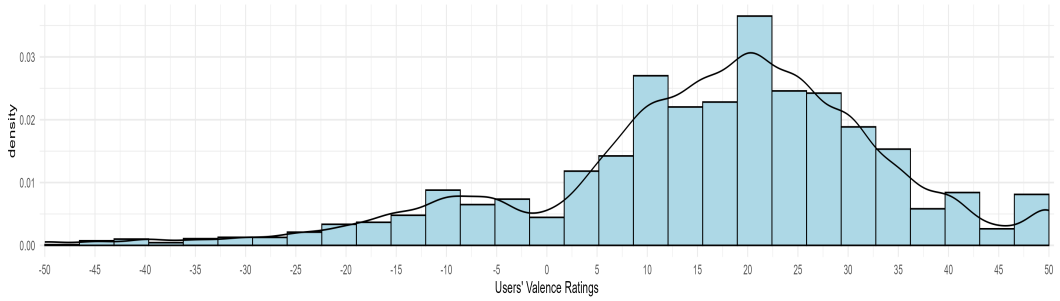

Figure 1: Distribution of users' self-reported valence ratings

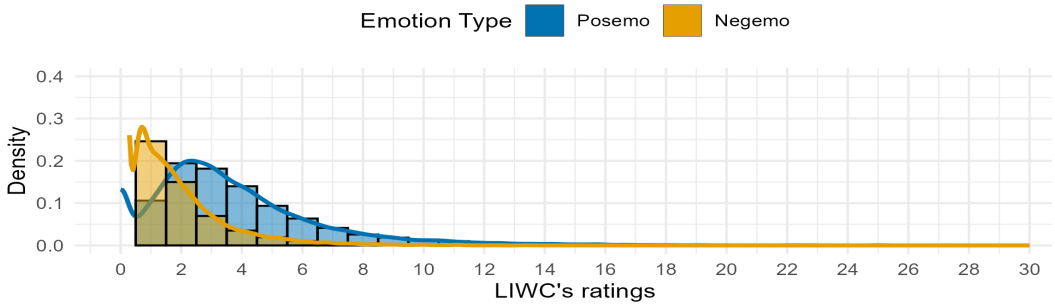

Figure 2: Distribution of LIWC's scores

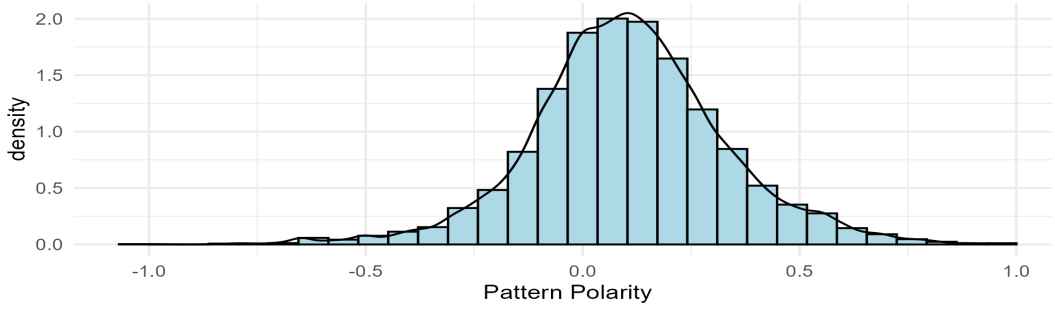

Figure 3: Distribution of Pattern.nl's polarity

A.4 Zero-shot setting results (All participants)

In the zero-shot condition, as discussed previously in the Results section, the models did not return values for all the texts, unlike LIWC and Pattern. ChocoLlama-8B-Instruct returned values for only 17378 texts, GEITje-7B-ultra for 9445 texts, and Reynaerde-7B-Chat for 446 estimates only. Pattern.nl turned out to be performing better on most of the texts than these models.

Polyserial correlation analyses revealed significant relationships between user ratings and the outputs of all three language models tested, with all user-model comparisons yielding p-values < 0.001. However, correlations between Reynaerde and traditional sentiment tools varied: LIWC15 (posemo) showed no significant association ($p = 0.12$), while LIWC15 (negemo) approached significance ($p < 0.05$) but did not meet the threshold for statistical reliability. In contrast, Pattern.nl demonstrated a modest yet statistically significant correlation with Reynaerde ($p = 0.022$). GEITje showed strong, statistically significant correlations with all lexicon-based models-LIWC15 (posemo and negemo) and

Pattern.nl (all $p < 0.001$), indicating robust agreement. Similarly, ChocoLlama's outputs correlated significantly with LIWC15 and Pattern.nl (all $p < 0.001$).

A.4.1 Distributional Analysis of the Models' Ratings

ChocoLlama-8B-Instruct showed a relatively balanced distribution compared to the other two LLMs (peaking around 3 and 5). GEITje-7B-ultra's ratings were skewed towards scores around 5. Reynaerde-7B-Chat's predictions had the narrowest distribution amongst all these three models.

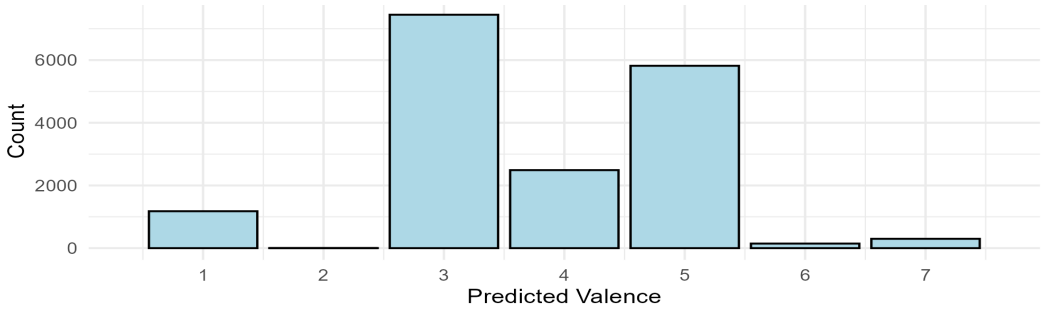

Figure 4: Distribution of Chocollama-8B-instruct's predicted valence scores

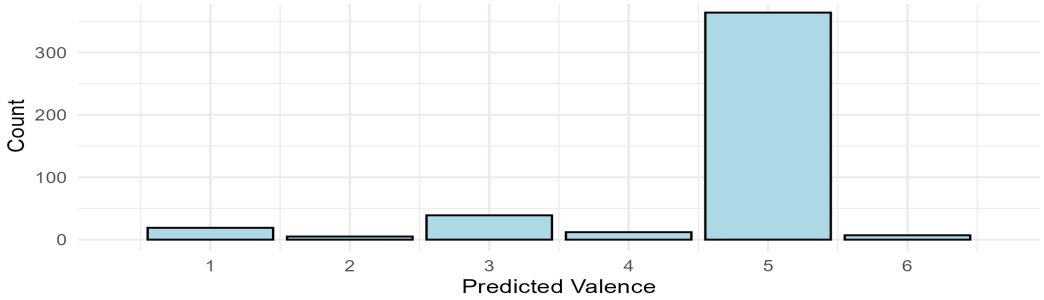

Figure 5: Distribution of Reynaerde-7B's valence scores

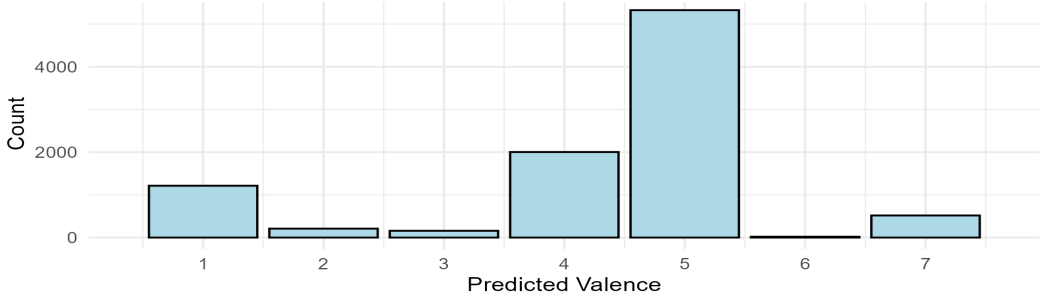

Figure 6: Distribution of GEITje-7B's valence scores

A.4.2 Box Plot Analysis of Users' Ratings vs. Model's Predictions

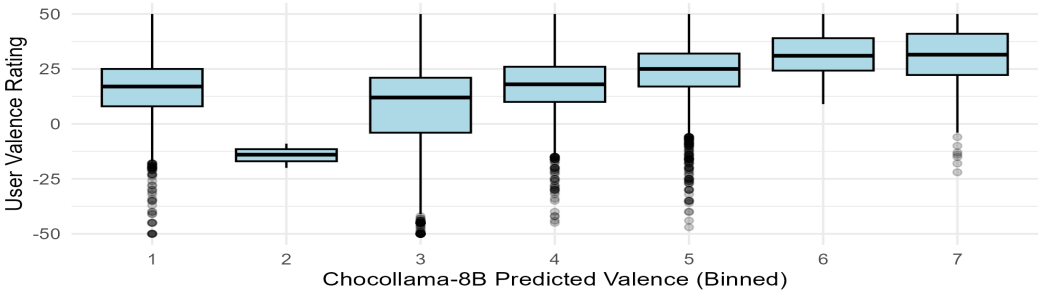

Figure 7: Chocollama-8B-instruct

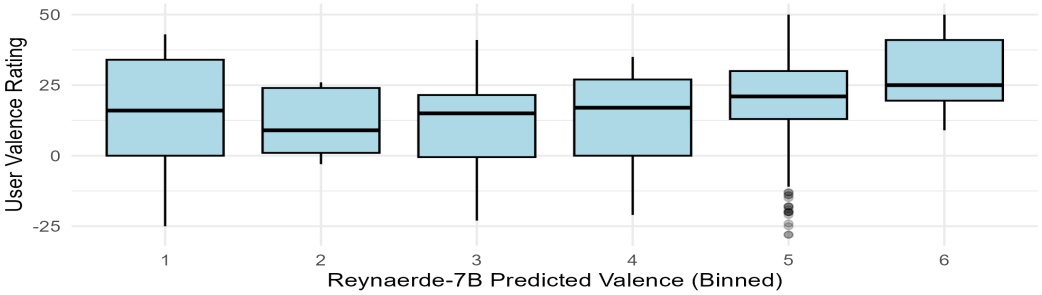

Figure 8: Reynaerde-7B

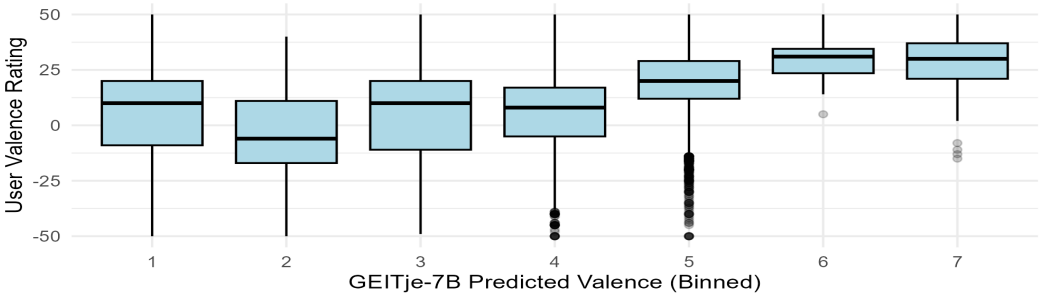

Figure 9: GEITje-7B

A.4.3 Scatter Plots

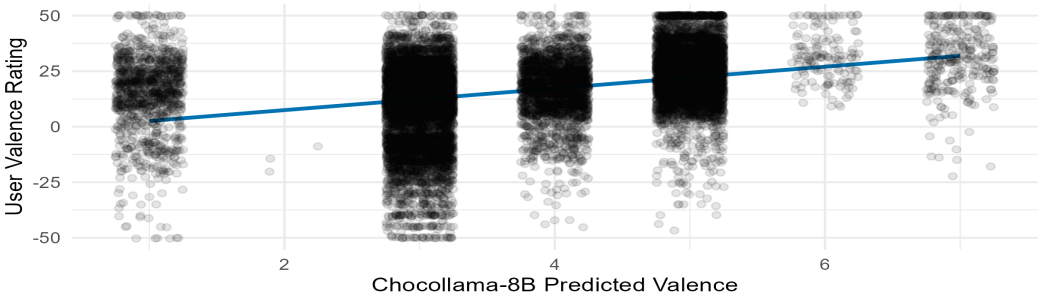

Figure 10: Chocollama-8B-instruct

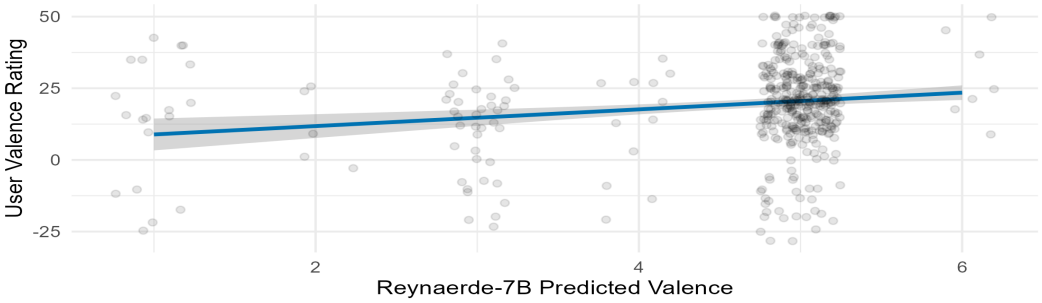

Figure 11: Reynaerde-7B

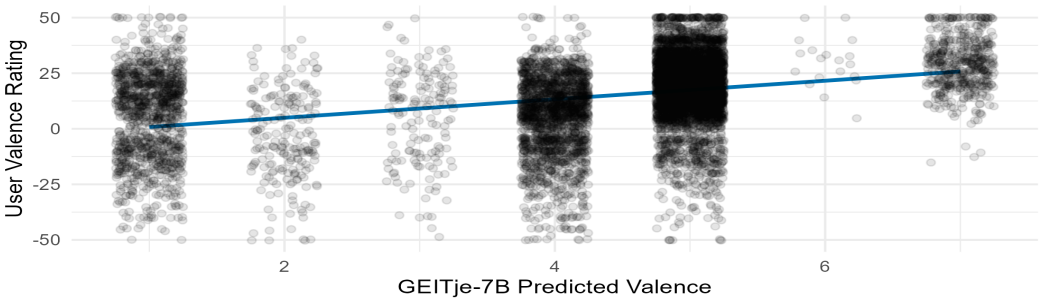

Figure 12: GEITje-7B

A.5 Few-shot setting results (All participants)

We also attempted a few-shot setting by providing five examples from the text along with the user ratings.

A.5.1 Prompt used

In this prompt, "(Text)" has been replaced with the actual responses of the participants.

"You are a Dutch language expert analyzing the valence of Belgian Dutch texts. Participants responded to the question:
"What is going on now or since the last prompt, and how do you feel about it?"
They also rated their own emotional valence on a continuous scale from -50 (very negative) to +50 (very positive). Your task is to read each response: text and rate its sentiment from 1 (very negative) to 7 (very positive). Return ONLY a single numerical rating enclosed in square brackets, e.g. [X]. Provide no explanation or additional text. Output Format: [number].

Below are a few examples of an input and the participant's valence rating to guide your rating:

Input: (Text1)
Output: [10]

Input: (Text2)
Output: [-10]

Input: (Text3)
Output: [30]

Input: (Text4)
Output: [45]

Input: (Text5)
Output: [40]"

We observed a similar trend in the LLMs chosen returned values for only a portion of the texts. Pattern.nl proved to be more effective than LIWC.

Table 3: Few-shot coverage and correlation coefficients across models and benchmarks wrt users

Model	**Coverage (N / %)**	**Pearson** r	**Polyserial** r
LIWC15 (posemo)	24,848 / 99.9%	0.21	0.23
LIWC15 (negemo)	24,848 / 99.9%	-0.23	-0.26
Pattern.nl	24,848 / 99.9%	0.31	0.33
Reynaerde-7B-Chat (English prompt)	7,219 / 29.0%	0.03	0.036
GEITje-7B-ultra (English prompt)	11,323 / 45.6%	-0.08	-0.09
Chocollama-8B-instruct(English prompt)	5,266 / 21.2%	-0.03	-0.029

In the few-shot setting, we provided five example texts with user ratings. Significance values (p-values) from polyserial correlation analyses were as follows:

A.5.2 Polyserial Correlation Between the models

Table 4: Polyserial correlations and significance for model-model comparisons

Comparison	**r**	**p**	**Significance**
LIWC(posemo) vs. Reynaerde-7B-Chat	0.11	< 0.001	significant
LIWC(negemo) vs. Reynaerde-7B-Chat	-0.004	0.916	non-significant
Pattern vs. Reynaerde-7B-Chat	0.044	0.001	significant
LIWC(posemo) vs. GEITje-7B-ultra	0.068	0.003	significant
LIWC(negemo) vs. GEITje-7B-ultra	0.131	< 0.001	significant
Pattern vs. GEITje-7B-ultra	-0.05	0.700	non-significant
LIWC(posemo) vs. Chocollama-8B-instruct	-0.052	0.479	non-significant
LIWC(negemo) vs. Chocollama-8B-instruct	-0.004	0.001	significant
Pattern vs. Chocollama-8B-instruct	-0.046	0.940	non-significant

A.5.3 Distributional Analysis of the Models' Ratings

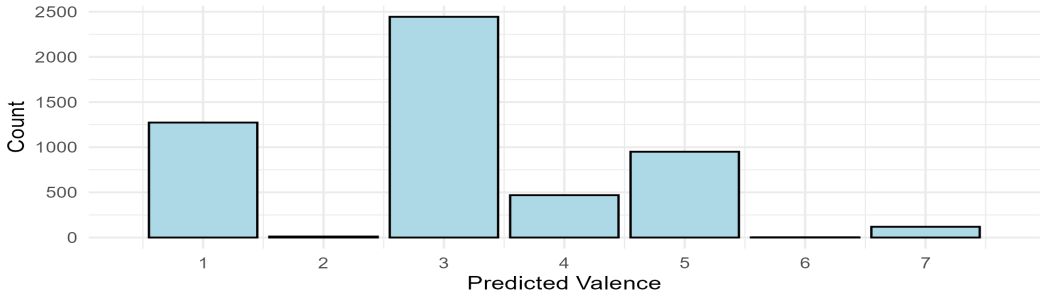

Figure 13: Distribution of Chocollama-8B-instruct's predicted valence scores

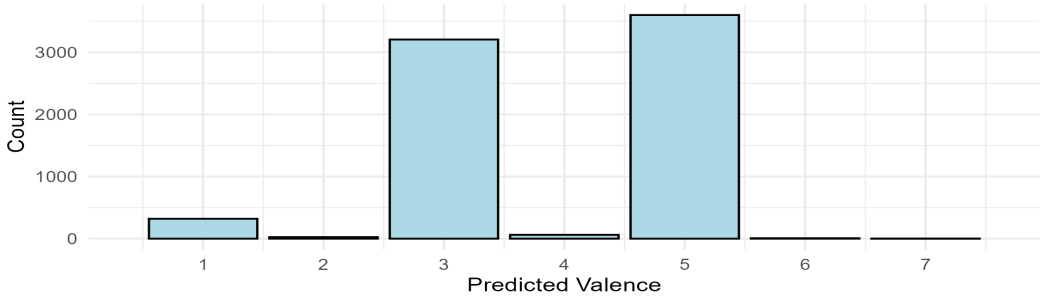

Figure 14: Distribution of Reynaerde-7B's valence scores

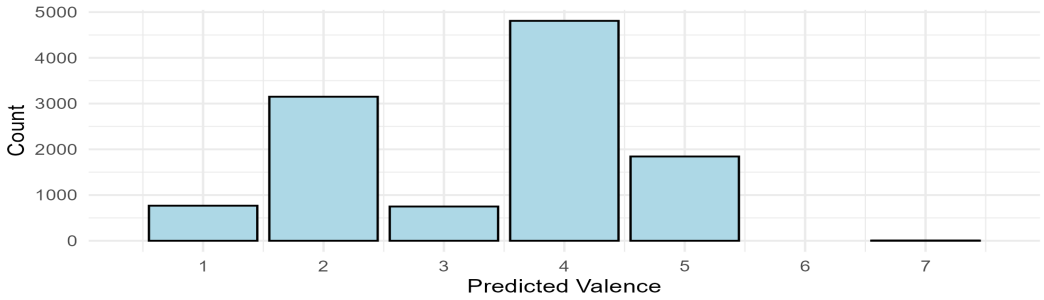

Figure 15: Distribution of GEITje-7B's valence scores

A.5.4 Box Plots of Users' Ratings vs. Model's Predictions

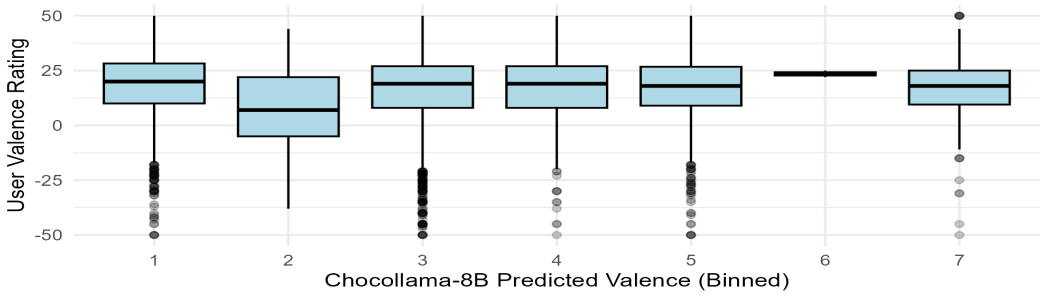

Figure 16: Chocollama-8B-instruct

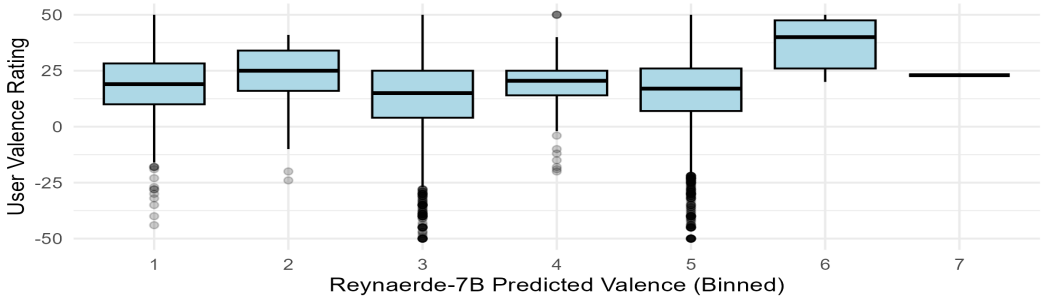

Figure 17: Reynaerde-7B

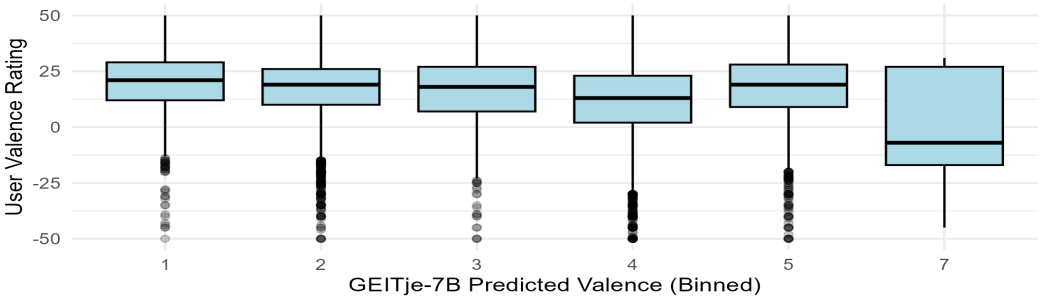

Figure 18: GEITje-7B

A.5.5 Scatter Plots

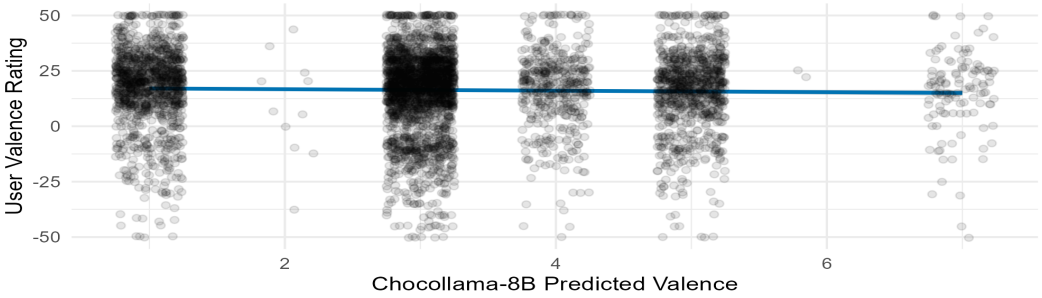

Figure 19: Chocollama-8B-instruct

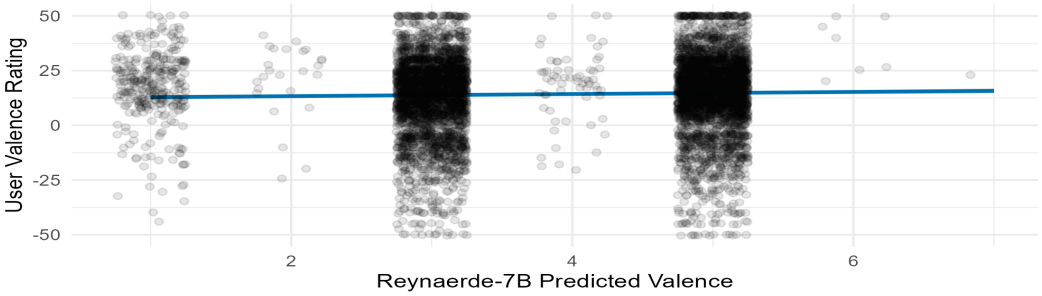

Figure 20: Reynaerde-7B

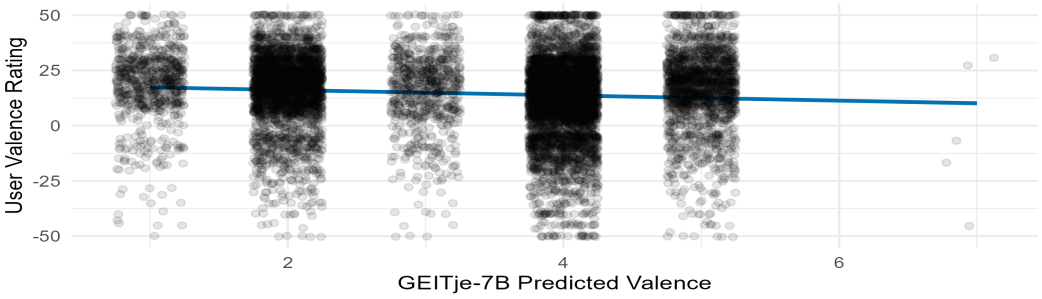

Figure 21: GEITje-7B

A.6 References

Mestdagh, M., Verdonck, S., Piot, M., Niemeijer, K., Kilani, G., Tuerlinckx, F., Kuppens, P., and Dejonckheere, E. (2023). m-Path: an easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. Frontiers in Digital Health, 5. https://doi.org/10.3389/fdgth.2023.1182175

Tamm, B., Poncelet, J., Barberis, M., Vandermosten, M., and Van hamme, H. (2024). Weakly supervised training improves Flemish ASR of non-standard speech. Frontiers in Digital Health, 5. https://doi.org/10.3389/fdgth.2023.1182175

A Technical Appendices and Supplementary Material

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction clearly state that we compare Dutch-tuned LLMs and lexicon-based tools on Flemish valence estimation, outline our dataset and methodology, and summarize the principal findings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We include a dedicated “Limitations” section that addresses model scope, dataset representativeness, coverage bias, psychometric constraints, and the empathy gap (Section 9).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer:[NA]

   Justification: This work is purely empirical and does not present new theoretical results or formal proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: We describe dataset composition, preprocessing steps, model prompting, correlation metrics, and statistical tests in detail,

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The anonymized text responses will be made available upon request to interested candidates, and the analysis and plotting code will be released once the paper is accepted.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail the exact prompt templates, data split by model coverage, correlation computation methods, and software versions

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report paired Fisher's z–transform tests with t-statistics and p-values for all key model comparisons.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

Justification: We did not specify GPU/CPU types or wall-clock runtimes for inference; we will include this information in the final version.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our study follows NeurIPS ethical guidelines for data collection and model evaluation, with no harmful practices or dual-use concerns.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We address how improved valence analysis can benefit psychological research and note risks such as sentiment profiling and privacy issues (Discussion).

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We use only de-identified, self-reported text data and publicly available models; no additional safeguards were required.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the LIWC 2015 license, Pattern.nl documentation, and the respective open-source licenses for the LLMs

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:We do not release new datasets or models; we analyze existing text responses and pretrained LLMs.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: **[TODO]**

    Justification: We did not include the full IRB protocol or compensation details in the main manuscript; we will add these to the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer:[No]

    Justification: IRB approval and risk disclosures are not currently described; we will include this information in the final version.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We detail the three Dutch-tuned LLMs, their prompting strategies, and coverage behaviors as central to our methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.