“An bhfuil Gaeilge agat?”: Differences in User Interaction and Assistant Responses Across Languages of European Origin in Large-scale Conversational Datasets

Aldan Creo[1]

[1]*Independent author, Dublin, IE*

Abstract

This study presents a comprehensive analysis of user interactions and assistant responses across 28 European languages using the WildChat and LMSYS datasets, addressing an existing gap in the understanding of multilingual conversational AI. We examine five specific dimensions: the topics discussed, the length of the conversations, the sentiment expressed, the toxicity of the interactions, and the quality of the responses. Our findings indicate significant cross-linguistic variations that have significant implications for the development and deployment of language models. Topic analysis shows a high degree of overlap across languages, indicating that users engage with similar subjects regardless of their linguistic background. We observe a positive correlation between the frequency of language use and conversation length, which suggests that different engagement patterns may be at play across language communities. Sentiment analysis indicates a high degree of consistency in neutral tones across languages, whereas toxicity levels vary considerably, with some languages exhibiting notably elevated scores. To assess response quality, we introduce a custom neural architecture based on the classification of user-assistant interaction triples. Our model achieved an accuracy of 0.82 and served to uncover variations in user satisfaction across language groups. Speakers of Romance languages exhibited higher levels of satisfaction, whereas those of Eastern European languages tended to show lower satisfaction with their interactions with the assistant. Our findings underscore the need for language-specific strategies in conversational AI development, particularly in content moderation, conversation design, and quality assessment. By highlighting the differences and commonalities in conversational interactions across languages, our work provides insights for researchers and developers seeking to better understand and address the needs of users across a diverse linguistic landscape.

Keywords

Artificial Intelligence, Natural Language Processing, Conversational AI, Multilingualism

1. Introduction

The development of conversational assistants has seen significant advancements in recent years, following the advent of Transformer-based architectures [1], as exemplified by the GPT family of models. These assistants are capable of engaging in conversations with users on a wide range of topics, providing information, answering questions, and even engaging in small talk [2].

The popularity of conversational assistants has therefore surged across various applications, including customer service and language learning [3, 4]. Notably, ChatGPT exceeded 100 million monthly active users within its first two months, a testament to the rising interest in conversational AI [5]. However, despite this booming interest, research and development in the field remain predominantly English-centric, with limited focus on other languages [6, 7].

A branch of research that is key to the development of conversational assistants is the analysis of user interactions and assistant responses in a large scale. Two main works in this area due to their size and diversity are the WildChat and LMSYS datasets [8, 9], which contain a large number of user interactions with conversational assistants and are publicly available. However, there is still work to be done in analyzing these datasets, due to their size and complexity.

This study aims to explore user interactions and assistant responses, focusing on linguistic differences

AICS’24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09–10, 2024, Dublin, Ireland
research@acmc.fyi (A. Creo)
https://acmc.fyi/ (A. Creo)
0000-0002-7401-5198 (A. Creo)

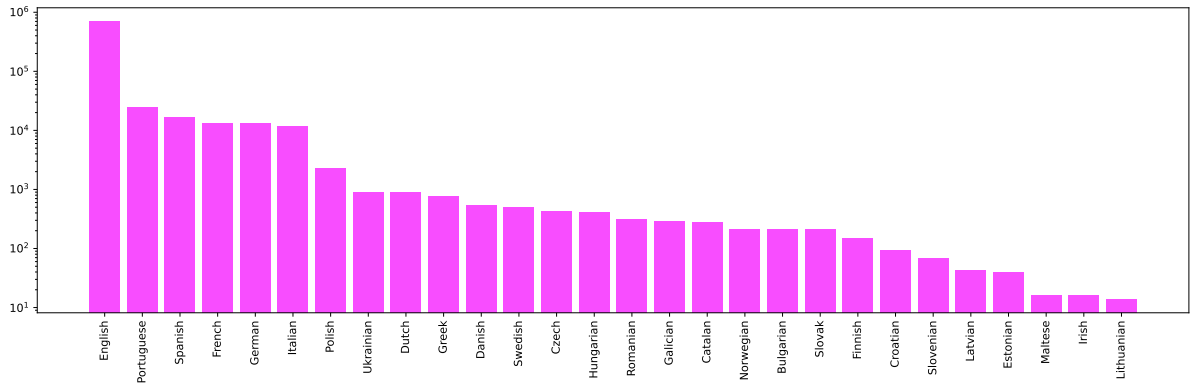

Figure 1: Number of examples per language.

across European languages. Previous research has only examined datasets as a whole [10], without delving into cross-language discrepancies—a gap our study seeks to fill. In contrast, we seek to provide insights into cross-language differences within large-scale conversational datasets, in order to inform the development of assistants that are better tailored to the needs of users of a wider range of languages. Our analysis focuses on language groups rather than the countries in which users reside. We acknowledge that languages like Spanish or Portuguese may have significant representation from Latin American users, contributing to the diversity of our findings.

The rest of this paper is structured as follows. The subsequent sections examine whether notable differences exist across languages in terms of **topics** discussed (**RQ1**), **length** of interactions (**RQ2**), **sentiment** expressed (**RQ3**), **toxicity** (**RQ4**), and **quality** (user satisfaction) of responses (**RQ5**). We then proceed to an integrative discussion and conclusion.

2. Datasets

In this section, we describe the datasets used for our analysis.

We combined the WildChat and LMSYS datasets [8, 9], which contain 990,372 and 1,000,000 examples, respectively. Nevertheless, the count of examples we analyze is lower, as we applied several preprocessing steps to clean the data. First, we discarded examples where the user's initial message contained fewer than five words, as short inputs hinder the accuracy of language detection. For instance, messages like "Test" were often misclassified, such as being incorrectly identified as Estonian. Furthermore, we excluded interactions where the toxicity scores were not defined, as these annotations are essential for our toxicity analysis.

To ensure a comprehensive representation of European languages, we included 28 languages, shown in Figure 1, together with the number of examples in each. However, we excluded languages such as Basque and Breton because of the lack of a sufficient number of examples in the datasets, as well as the lack of language models trained specifically for these languages, which could introduce bias when using them in our analysis. After applying these filtering criteria, we obtained a total of 781,376 examples.

3. RQ1: Topics

In this section, we analyze the distribution of topics across languages. We utilize language tags as a clustering feature for semantically-informed embeddings. We hypothesize that if speakers of different languages tend to discuss distinct topics, language tags should form clusters with clear boundaries. However, if topics are more uniformly distributed across languages, the boundaries will be blurrier.

We commence our analysis by selecting all initial user messages in the dataset, as these are the most representative, establishing the context for subsequent discourse. We then use two multilingual models

Table 1
Mean number of messages, user words, and assistant words per conversation.

Language	Messages	User words	Assistant words
Bulgarian	3.38	92.68	158.92
Catalan	4.29	98.00	254.60
Croatian	5.78	103.07	248.80
Czech	3.76	45.89	176.18
Danish	3.74	83.94	219.82
Dutch	3.72	88.88	217.99
English	3.68	89.28	264.70
Estonian	3.38	44.26	100.77
Finnish	3.96	57.18	211.57
French	4.24	92.68	298.27
Galician	4.71	44.84	340.51
German	4.31	87.44	272.59
Greek	3.44	49.90	121.65
Hungarian	3.19	46.03	128.22
Irish	3.25	33.69	193.06
Italian	4.77	66.66	312.91
Latvian	4.29	61.05	186.12
Lithuanian	2.14	7.14	88.07
Maltese	4.00	16.75	163.88
Norwegian	3.44	51.00	219.81
Polish	3.83	59.40	171.79
Portuguese	4.78	83.19	319.67
Romanian	3.83	76.29	221.79
Slovak	4.16	67.23	227.99
Slovenian	3.07	60.65	153.43
Spanish	4.59	92.25	307.73
Swedish	4.89	100.37	253.91
Ukrainian	5.72	109.19	250.05

from the Sentence Transformers library [11]—`paraphrase-multilingual-mpnet-base-v2` and `paraphrase-multilingual-MiniLM-L12-v2`—to generate embeddings of these messages.

To evaluate how well messages cluster based on language, we conduct a silhouette score analysis (ranging from -1 to 1), which measures cluster cohesion and separation. A score of 1 indicates well-defined, separated clusters, while a score of 0 suggests overlap. Negative scores indicate poorly defined or incorrect clusters [12].

Due to the high computational cost of this analysis, which scales quadratically with the number of examples, we performed the silhouette score calculation on 20 randomly-selected subsets, each representing 5% of the dataset, and then averaged the results. The mean silhouette scores for the two models were $-0.121(24)$ — mean(s.d.) — and $-0.115(30)$. These negative scores indicate significant semantic overlap between clusters, confirming that languages are not effective clustering tags.

Per-language silhouette scores, which align with the overall results, are provided in the supplementary material. Based on this analysis, we conclude that users across linguistic groups engage with similar themes, rather than showing strong language-specific patterns. This finding highlights the universality of topics and suggests that conversational assistants must prioritize general topic coverage and flexibility. Moreover, the lack of clustering implies that cultural or regional nuances may play a smaller role in topic differentiation than previously expected.

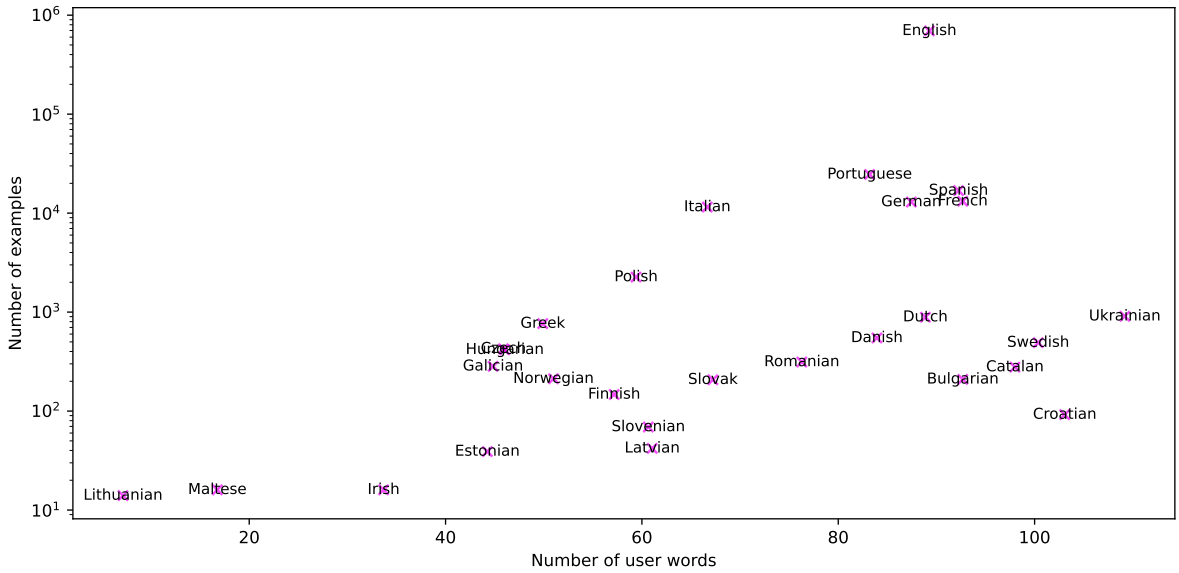

Figure 2: Scatter plot of the mean number of user words and the number of examples, for each language.

4. RQ2: Length

In this section, we explore the correlation between the mean number of words written by users in a conversation and the number of examples per language.

As both variables are not normally distributed, we study their interaction by calculating the Spearman Rank correlation coefficient, which ranges from -1 to 1. A value of -1 indicates a perfect negative correlation; 0, no correlation; and 1 signifies a perfect positive correlation. The coefficient evaluates to 0.51, indicating a moderate positive correlation between the number of user words and the number of examples in a language. With a p-value of $0.0055 < \alpha = 0.05$, the correlation is statistically significant.

While not a perfect correlation, the strength of the relationship suggests that the number of user words can be considered a reasonably reliable predictor of the number of examples in a language. Figure 2 presents a scatter plot illustrating this relationship. Additionally, Table 1 provides the per-conversation mean number of messages, user words, and assistant words for each language.

In essence, we find that speakers of languages with more interactions in conversational datasets tend to engage in longer conversations, as measured by the number of user words. This trend provides moderate support for the notion that there exist behavioral differences across groups of language users, which may be indicative of cultural or linguistic factors influencing conversation length. However, further research is warranted to ascertain the underlying reasons for these differences.

5. RQ3: Sentiment

In this section, we explore sentiment differences across languages, as it allows to understand the emotional tone of conversations and provides insights into both user experience and interaction quality.

We use the pre-trained multilingual classification model `twitter-XLM-roBERTa-base` [13] for our analysis. This model is particularly suited to our multilingual setting, as it has been trained on a large, diverse corpus of Twitter messages in various languages.

Our analysis focuses on both the user and the assistant first messages in each conversation, as these set the tone and are likely the most representative of the overall sentiment. We calculate sentiment scores for both messages, then aggregate these scores across languages. We show the results in Table 2.

One clear observation is that the assistant's messages tend to be more positive than the users', with an overall mean of $0.2205(15)$ compared to $0.16(11)$. This difference likely reflects the Reinforcement

Table 2
Mean and standard deviation for the sentiment scores of the first user (U) and assistant (A) messages.

Language	Negative (U)	Neutral (U)	Positive (U)	Negative (A)	Neutral (A)	Positive (A)
Bulgarian	0.133(85)	0.74(12)	0.128(93)	0.18(16)	0.63(19)	0.19(17)
Catalan	0.27(20)	0.60(19)	0.13(11)	0.28(19)	0.51(18)	0.2102(18)
Croatian	0.30(20)	0.55(20)	0.15(13)	0.26(19)	0.53(19)	0.21(13)
Czech	0.25(18)	0.61(20)	0.14(11)	0.23(16)	0.57(18)	0.200(15)
Danish	0.21(16)	0.63(17)	0.16(12)	0.25(18)	0.52(17)	0.22(15)
Dutch	0.24(21)	0.64(22)	0.13(11)	0.25(20)	0.569(21)	0.18(17)
English	0.26(20)	0.58(19)	0.16(11)	0.28(20)	0.49(17)	0.22(15)
Estonian	0.31(14)	0.45(18)	0.24(12)	0.24(11)	0.45(18)	0.306(20)
Finnish	0.34(25)	0.51(22)	0.16(14)	0.30(20)	0.483(17)	0.22(16)
French	0.27(21)	0.481(20)	0.25(17)	0.28(19)	0.410(17)	0.311(18)
Galician	0.20(19)	0.68(19)	0.121(92)	0.23(15)	0.56(16)	0.21(14)
German	0.22(18)	0.68(20)	0.11(12)	0.2150(17)	0.63(20)	0.151(17)
Greek	0.26(15)	0.55(16)	0.19(13)	0.24(14)	0.53(14)	0.24(14)
Hungarian	0.20(14)	0.66(18)	0.14(12)	0.20(18)	0.60(20)	0.20(16)
Irish	0.284(18)	0.53(21)	0.19(18)	0.372(31)	0.394(23)	0.23(25)
Italian	0.28(23)	0.59(23)	0.14(14)	0.28(22)	0.53(21)	0.19(17)
Latvian	0.22(15)	0.67(14)	0.106(86)	0.28(22)	0.51(21)	0.20(17)
Lithuanian	0.22(11)	0.71(11)	0.070(30)	0.244(79)	0.59(13)	0.165(84)
Maltese	0.26(16)	0.58(22)	0.16(12)	0.16(11)	0.58(17)	0.25(18)
Norwegian	0.194(17)	0.65(19)	0.15(14)	0.220(18)	0.51(18)	0.27(19)
Polish	0.19(17)	0.70(19)	0.116(11)	0.193(18)	0.63(20)	0.18(16)
Portuguese	0.190(17)	0.67(18)	0.142(13)	0.20(16)	0.60(17)	0.19(15)
Romanian	0.21(18)	0.66(19)	0.13(12)	0.196(17)	0.60(19)	0.21(17)
Slovak	0.25119(19)	0.60(21)	0.15(13)	0.22(15)	0.58(17)	0.20(15)
Slovenian	0.18(17)	0.67(19)	0.14(10)	0.228(16)	0.55(19)	0.217(14)
Spanish	0.21(19)	0.65(20)	0.14(13)	0.22(17)	0.56(17)	0.22(16)
Swedish	0.25(19)	0.58(20)	0.17(16)	0.24(19)	0.497(19)	0.262(20)
Ukrainian	0.17(12)	0.69(13)	0.138(87)	0.19(14)	0.64(16)	0.17(12)

Learning from Human Feedback (RLHF) paradigm used for training [14], which encourages the assistant to maintain a more positive and helpful tone. Interestingly, the assistant also shows a slightly higher negative sentiment score ($0.28(20)$) than the users ($0.26(20)$). This could also be attributed to RLHF, which prompts it to refrain from engaging in potentially toxic conversations, thus increasing the frequency of negative sentiment classifications in those contexts [15]. Overall, however, both user and assistant messages tend to be neutral, with a mean neutral score of $0.50(18)$ and $0.58(19)$, respectively.

These observations show that while there are differences in the sentiment expressed by users and assistants, the sentiment across languages tends to remain fairly neutral.

6. RQ4: Toxicity

We perform a toxicity analysis to identify potential differences, for every message in the dataset (a total of $2{,}951{,}678$). There exist 11 types of toxicity annotations generated by the dataset creators with the OpenAI Moderations API. To simplify our analysis, we aggregate these categories into two general toxicity scores using the mean and maximum values across the categories. We do this for each example. Then, we calculate the mean of the two across the examples, which we present in Table 3.

In the rest of this section, we focus on the perspective of the user messages, as they are generally causative of the toxicity in the assistant's side of the conversation. To identify significant differences in toxicity across languages, we employ the Kruskal-Wallis test. This is a non-parametric method well-suited to compare medians of toxicity scores across multiple independent samples, especially given the non-normal distribution of toxicity values, which cluster near the extremes. We test a null

Table 3
Mean and standard deviation of averaged and maximum toxicity scores of user (U) and assistant (A) messages.

Language	Average (U)	Maximum (U)	Average (A)	Maximum (A)
Bulgarian	0.0009(53)	0.0086(55)	0.00037(82)	0.0028(73)
Catalan	0.006(26)	0.05(18)	0.006(25)	0.04(17)
Croatian	0.0010(75)	0.007(44)	0.0005(20)	0.004(17)
Czech	0.002(12)	0.0143(74)	0.0015(80)	0.014(77)
Danish	0.003(16)	0.03(12)	0.003(17)	0.02(11)
Dutch	0.006(21)	0.05(18)	0.006(26)	0.042(17)
English	0.006(25)	0.05(18)	0.008(31)	0.06(22)
Estonian	0.00042(64)	0.0026(46)	0.0006(12)	0.005(12)
Finnish	0.00084(31)	0.007(27)	0.0009(53)	0.007(54)
French	0.005(19)	0.04(15)	0.005(22)	0.041(17)
Galician	0.0016(67)	0.014(68)	0.0008(40)	0.0073(35)
German	0.006(27)	0.04(16)	0.008(32)	0.06(20)
Greek	0.0007(19)	0.0046(14)	0.0004(11)	0.0030(92)
Hungarian	0.0014(58)	0.012(57)	0.0012(98)	0.008(52)
Irish	0.00032(45)	0.00222(41)	0.00042(57)	0.0025(39)
Italian	0.003(14)	0.02(11)	0.003(17)	0.03(14)
Latvian	0.002(11)	0.014(85)	0.0004(12)	0.004(13)
Lithuanian	0.00010(3)	0.00039(0)	0.00003(3)	0.00029(37)
Maltese	0.0005(11)	0.005(12)	0.0006(18)	0.006(19)
Norwegian	0.0007(24)	0.0066(25)	0.0006(20)	0.006(20)
Polish	0.002(12)	0.02(10)	0.002(15)	0.02(11)
Portuguese	0.002(13)	0.017(91)	0.002(12)	0.0133(85)
Romanian	0.0012(75)	0.011(68)	0.0008(56)	0.007(54)
Slovak	0.0020(72)	0.019(74)	0.0011(65)	0.011(72)
Slovenian	0.003(17)	0.02(14)	0.0008(26)	0.007(23)
Spanish	0.003(14)	0.02(11)	0.003(18)	0.02(12)
Swedish	0.005(15)	0.05(15)	0.003(10)	0.0262(11)
Ukrainian	0.0007(64)	0.006(45)	0.0005(44)	0.004(30)

hypothesis H_0 = "The medians of the toxicity scores are equal across languages" with $\alpha = 0.05$.

Since toxicity scores are continuous and generally skewed towards values close to 0, directly applying the Kruskal-Wallis test might exaggerate differences between languages due to minor deviations in low-toxicity messages. To mitigate this, we round the scores to two decimal places, reducing the number of unique values and treating very close values as ties. The outcomes of the Kruskal-Wallis test, as well as independent analyses for each toxicity category and aggregated scores, are presented in Table 4.

With the exception of the "self-harm" category, the p-values for all other categories are less than α, indicating significant differences in toxicity across languages. This necessitates pairwise comparisons to determine which languages exhibit meaningful disparities. We use Dunn's test for this purpose, with the null hypothesis H_0 = "The probability that a randomly selected message from one language has a higher toxicity score than one from another language is 0.5." To control for the increased risk of Type I errors due to multiple comparisons, we apply the Bonferroni correction to the p-values. The results of these pairwise comparisons for the averaged toxicity scores are shown in Figure 3, while results for other configurations are included in the supplementary materials.

Key findings from the pairwise comparisons include:

1. *No Statistically Significant Differences*: For most languages—such as Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Romanian, Slovak, and Slovenian—the comparisons are not statistically significant. In these cases, we cannot reject H_0, indicating it is equally likely that a randomly selected message from one of these languages has a higher or lower toxicity score than another.
2. *Significant Differences in Specific Languages*: Pairwise comparisons involving Dutch, English,

Table 4
Results of the Kruskal-Wallis test for user messages.

Toxicity category	H-statistic	p-value
harassment	3091.9	0.0
harassment/threatening	163.4	1.8×10^{-21}
hate	8424.2	0.0
hate/threatening	314.2	1.1×10^{-50}
self-harm	811.6	1.4×10^{-153}
self-harm/instructions	30.2	3.1×10^{-1}
self-harm/intent	144.1	5.9×10^{-18}
sexual	717.7	7.3×10^{-134}
sexual/minors	730.9	1.2×10^{-136}
violence	1774.8	0.0
violence/graphic	666.0	5.0×10^{-123}
max_toxicity_score	3511.0	0.0
avg_toxicity_score	2307.0	0.0

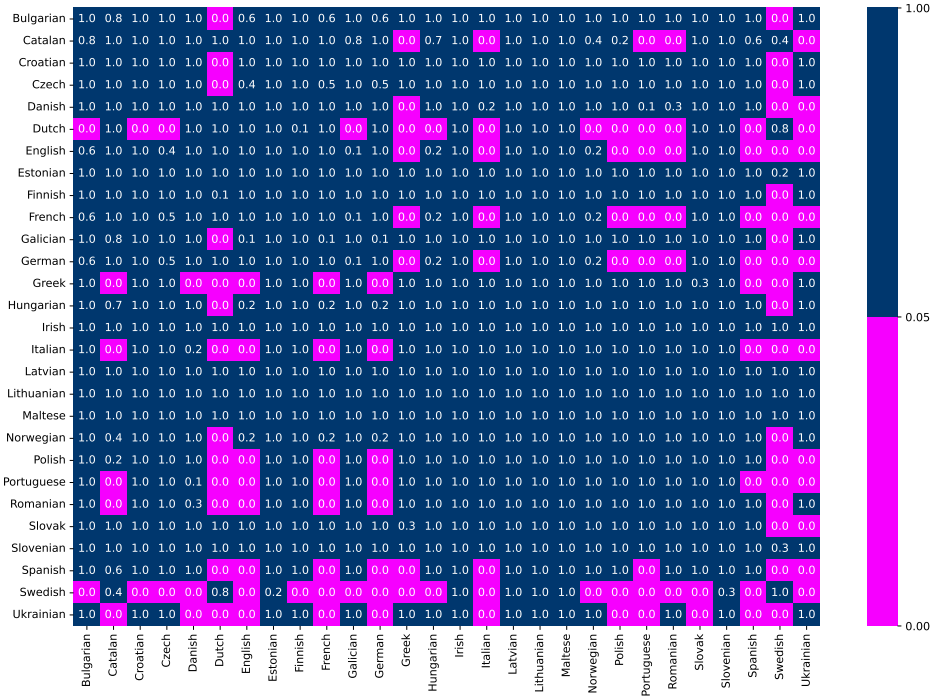

Figure 3: Dunn's test p-values for the user's mean toxicity score. We only color significant cells.

French, German, Spanish, Swedish, and Ukrainian reveal significant differences compared to several languages from the first group (Greek, Italian, Polish, Portuguese, Romanian, Spanish, Swedish and Ukrainian). These languages show distinct toxicity profiles, although notable exceptions exist within this subset.

3. *Group of Similar Toxicity*: While Dutch, English, French, and German exhibit significant differences compared to other languages, they are not significantly different from one another, suggesting a group of languages with similar, higher-than-average toxicity scores.
4. *Highest Average Toxicity*: Languages such as Catalan, Dutch, English, German, and Swedish display the highest average toxicity scores, with Dutch, English, German, and Swedish identified

Table 5
Search space for the hyperparameters of the classification head, and best values found.

Hyperparameter	Search space	Best value
Learning rate	$[1 \times 10^{-6}, 1 \times 10^{-3}] \in \mathbb{R}$	4.7×10^{-4}
Batch size	$\{1, 4, 8, 16, 32, 64, 128\}$	1
Hidden size	$[10, 3000] \in \mathbb{N}$	884
Number of layers	$[1, 5] \in \mathbb{N}$	4
Use bias	{True, False}	True
Apply weighted loss	{True, False}	False
Loss label smoothing	$[0.0, 0.3] \in \mathbb{R}$	0.2
Activation function	{Tanh, ReLU, LeakyReLU, Sigmoid, GELU, SiLU, Mish}	Mish

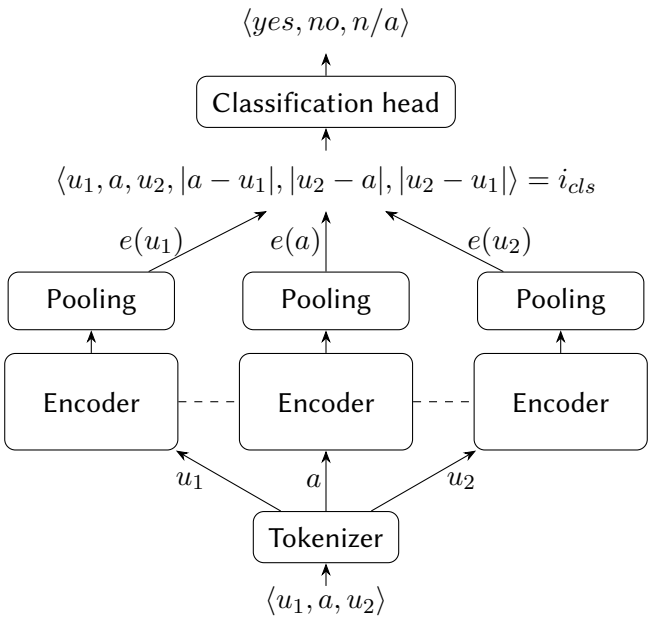

Figure 4: Architecture of our model to evaluate the quality of assistant responses.

as significantly different in the majority of pairwise comparisons.

Overall, the results highlight considerable variability in the toxicity of conversations across languages. The distribution of toxicity scores varies significantly between some languages, suggesting that conversational toxicity may be influenced by a range of factors, including cultural backgrounds, the structure of the language itself, potential biases in toxicity tagging, or a combination of these influences.

7. RQ5: Quality

In this section, we assess the quality of the assistant's responses, which we define as *how well the assistant meets the user's needs and expectations*. To the best of our knowledge, no large-scale multilingual dataset exists with labels for assistant response quality, such as those used in RLHF [16]. This motivated us to develop a custom architecture tailored for evaluating assistant responses in a multilingual setting.

Our approach is inspired by the siamese architecture proposed in Sentence-BERT [11]. While the original architecture encodes *two* inputs and trains a classification head that is later discarded to retain only the fine-tuned encoder, we focus on training the classification head with *three* inputs to evaluate the quality of assistant responses, the encoder remaining frozen. The classification task asks "Is the user satisfied with the assistant's response?", with classes "Yes," "No," and "N/A" (not applicable).

We structure the model to process $\langle u_1, a, u_2 \rangle$ triples, where u_1 is the user's initial message, a is the

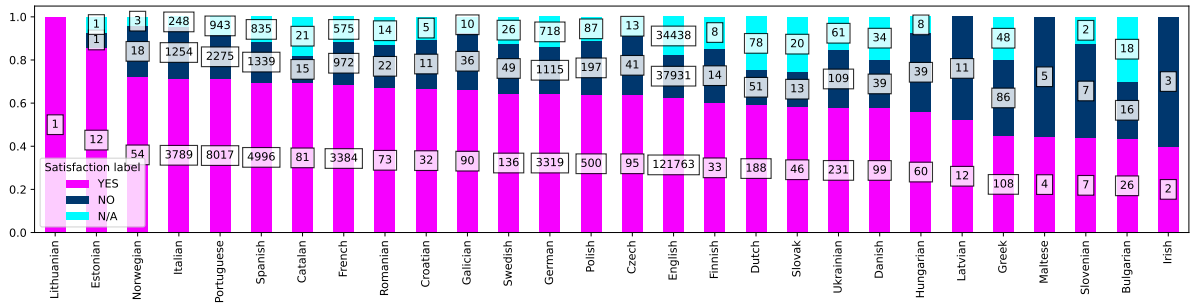

Figure 5: Percentage and total counts of satisfaction labels for each language.

assistant's response, and u_2 is the user's second message. The goal is to capture the semantics of the user-assistant exchange and assess whether the assistant's response satisfies the user's original query. For instance, if the user were to ask about the weather, the embedding of their first message may belong to a subspace "weather"; if the assistant were to respond with the weather forecast, the embedding of the assistant's response could also belong to the same subspace. The user, having received a satisfactory response, may then express their satisfaction in their second message, the embedding of which could encode a positive sentiment. We expect the model to be able to capture the distances between the embeddings of the messages (*e.g.*, the assistant's response addressing the user's first message), as well as the semantics of the messages themselves (*e.g.*, the user's satisfaction with the assistant's response).

Formally, our model takes a tokenized input of shape $3 \times L$, where L is the maximum length of the tokenized input (shorter sequences are padded), the three messages being stacked along a new dimension. The model outputs a pooled representation $e(i)$ for each message in the triple, where $i \in u_1, a, u_2$, producing embeddings $e(u_1)$, $e(a)$, and $e(u_2)$ of dimension $d_e = 768$.

$$\text{sentence representation} = e(i) \in \mathbb{R}^{d_e} \quad \text{for} \quad i \in \{u_1, a, u_2\} \tag{1}$$

We concatenate these embeddings with their absolute differences into a single vector i_{cls}:

$$i_{cls} = \text{concatenate}(e(u_1), e(a), e(u_2), |e(a) - e(u_1)|, |e(u_2) - e(a)|, |e(u_2) - e(u_1)|) \in \mathbb{R}^{6 \times e} \tag{2}$$

We pass i_{cls} through a fully connected feedforward neural network with four layers, W_n and b_n being the weights and biases of layer n, each with a hidden size of 884 and Mish [17] activation functions to obtain the output logits o_{cls}:

$$o_{cls} = W_4 \cdot \text{Mish}(W_3 \cdot \text{Mish}(W_2 \cdot \text{Mish}(W_1 \cdot i_{cls} + b_1) + b_2) + b_3) + b_4 \in \mathbb{R}^3 \tag{3}$$

Finally, we apply an argmax operation on o_{cls} to determine the class prediction. A diagram of this architecture is shown in Figure 4. We used Optuna [18] to optimize the hyperparameters of the classification head and training. Table 5 shows the search space and best values found.

For training, we built a dataset of $\langle u_1, a, u_2 \rangle$ triples by sampling up to 1000 conversations per language with at least three messages. We manually annotated 1000 examples using the Argilla platform [19]. We also generated translations for non-English conversations to assist the annotation process using the `EuroLLM-1.7B` model [20]. We did not assess the quality of the translations, as they were utilized solely for contextual purposes, thereby facilitating a general comprehension of the discourse. Furthermore, we established explicit guidelines to facilitate a uniform interpretation of the task. Our annotation process maintained reasonable balance across languages, with an average of $37(33)$ annotations/language. Specific counts and annotation guidelines are provided in the supplementary material.

We trained the model using cross-entropy loss, optimized with AdamW [21] for 10 epochs, with the default hyperparameters in the Hugging Face Transformers library [22], $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. For the encoder, we utilized the `paraphrase-multilingual-mpnet-base-v2` pretrained weights [11]. We reserved 10% of the data for validation and another 10% for testing,

performing 10-fold cross-validation. The model achieved a mean test accuracy of $0.820(45)$, significantly outperforming both a random baseline ($0.339(42)$) and a majority class classifier ($0.594(44)$).

When applying the model to the full dataset (raw numbers are reported in the supplementary material), the results (Figure 5) reveal notable discrepancies in user satisfaction across linguistic groups. Disregarding languages with a limited number of examples, where the results may lack representativeness, we observe that the satisfaction of users with the assistant's responses is generally high and tends to be higher for users speaking languages belonging to the Romance language family [23], including Italian, Portuguese, Spanish, Catalan, French, and Romanian. In contrast, users speaking languages that originated in Eastern Europe, such as Bulgarian, Greek, and Hungarian, exhibit a lower level of satisfaction with the assistant's responses. English-speaking users demonstrate a satisfaction level that is comparable to the overall mean, similar to that observed in German and Polish, although the latter two languages exhibit a higher percentage of unsatisfied users.

8. Discussion

Our analysis of the WildChat and LMSYS datasets reveals significant differences in user-assistant interaction across European languages. These findings contribute to a more nuanced understanding of multilingual conversational AI and highlight the importance of considering linguistic diversity in the development and evaluation of language models.

The lack of clear clustering by language in our topic analysis (**RQ1**) suggests that users across different languages engage with conversational assistants on a wide variety topics. This finding is encouraging, as it indicates that the assistants are capable of handling a wide range of subjects across multiple languages. However, it also emphasizes the need for language models to be equally proficient in diverse topics across all supported languages.

Our analysis of conversation length (**RQ2**) revealed a positive correlation between the number of user words and the number of examples in a language. While further investigation is needed to understand the underlying factors driving this correlation, it suggests that users in languages with more examples may be more likely to engage in longer conversations, possibly due to a higher level of comfort or familiarity with the conversational AI system or a better performance of the assistant in those languages.

The sentiment analysis (**RQ3**) showed that both user and assistant messages tend to be neutral, with assistant responses generally being either more positive or negative. This is a consistent result across languages, and suggests that the current training approaches are effective in maintaining a coherent tone across different linguistic contexts.

Perhaps our most striking finding relates to toxicity (**RQ4**). The significant differences in toxicity levels across languages highlight the need for language-specific approaches to content moderation and toxicity detection. This is particularly important for languages like Dutch, English, German, and Swedish, which also exhibited higher average scores.

Finally, our quality analysis (**RQ5**) revealed considerable variations in user satisfaction across different language groups. The higher satisfaction levels among Romance language speakers and lower levels among Eastern European language speakers underscore the importance of tailoring conversational AI to specific linguistic and cultural contexts.

These findings collectively emphasize the need for a more nuanced, language-specific approach to the development and evaluation of conversational AI. While current models show promise in their ability to engage across multiple languages, there is still significant room for improvement in addressing language-specific challenges and user expectations.

9. Conclusion

This study presents the first comprehensive analysis of differences in user interaction and assistant responses across a wide range of European-origin languages for the WildChat and LMSYS datasets. Our work addresses a critical gap in existing literature, mainly treating these datasets as homogeneous.

By examining topics, conversation length, sentiment, toxicity, and response quality, we have uncovered significant variations across languages that have important implications for the development and deployment of conversational AI systems. Our findings highlight the need for more nuanced, language-specific approaches in areas such as content moderation and quality assessment.

The insights gained from this study are crucial for ensuring that the perspectives and needs of non-English speakers are adequately represented in the development of conversational AI. As the use of these systems continues to grow globally, it is imperative that they are designed to provide equitable and high-quality experiences across all languages.

Future work should focus on developing language-specific strategies for improving conversational AI, particularly in areas where we observed significant differences, such as toxicity levels and user satisfaction. Additionally, expanding this analysis to include non-European languages would provide a more comprehensive global perspective on multilingual conversational AI.

In conclusion, our work seeks to contribute to a more inclusive and effective approach to conversational AI development, providing insight into the importance of linguistic diversity in creating truly global and user-centric AI systems. We hope that these findings will inform and inspire future research and development efforts in multilingual conversational AI, ultimately leading to more equitable and effective language technologies for users worldwide.

Supplementary materials

Supplementary materials are available at https://github.com/ACMCMC/eur-langs-convs-analysis.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All you Need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[2] A. Casheekar, A. Lahiri, K. Rath, K. S. Prabhakar, K. Srinivasan, A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions, Computer Science Review 52 (2024) 100632. URL: https://www.sciencedirect.com/science/article/pii/S1574013724000169. doi:10.1016/j.cosrev.2024.100632.

[3] V. Katragadda, Automating Customer Support: A Study on the Efficacy of Machine Learning-Driven Chatbots and Virtual Assistants, IRE Journals 7 (2023) 600–610. URL: https://www.irejournals.com/.

[4] L. Kohnke, B. L. Moorhouse, D. Zou, ChatGPT for Language Teaching and Learning, RELC Journal 54 (2023) 537–550. URL: http://journals.sagepub.com/doi/10.1177/00336882231162868. doi:10.1177/00336882231162868.

[5] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, Y. Tang, A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development, IEEE/CAA Journal of Automatica Sinica 10 (2023) 1122–1136. URL: https://www.ieee-jas.net/en/article/doi/10.1109/JAS.2023.123618. doi:10.1109/JAS.2023.123618.

[6] E. Razumovskaia, G. Glavas, O. Majewska, E. M. Ponti, A. Korhonen, I. Vulic, Crossing the Conversational Chasm: A Primer on Natural Language Processing for Multilingual Task-Oriented Dialogue Systems, Journal of Artificial Intelligence Research 74 (2022) 1351–1402. URL: http://www.jair.org/index.php/jair/article/view/13083. doi:10.1613/jair.1.13083.

[7] S. Park, AI Chatbots and Linguistic Injustice, Journal of Universal Language 25 (2024) 99–119. URL: http://www.sejongjul.org/archive/view_article?doi=10.22425/jul.2024.25.1.99. doi:10.22425/jul.2024.25.1.99.

[8] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, Y. Deng, WildChat: 1M ChatGPT Interaction Logs in the Wild, 2023. URL: https://openreview.net/forum?id=Bl8u7ZRlbM.

[9] L. Zheng, W.-L. Chiang, Y. Sheng, T. Li, S. Zhuang, Z. Wu, Y. Zhuang, Z. Li, Z. Lin, E. Xing, J. E. Gonzalez, I. Stoica, H. Zhang, LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset, 2023. URL: https://openreview.net/forum?id=BOfDKxfwt0.

[10] Y. Deng, W. Zhao, J. Hessel, X. Ren, C. Cardie, Y. Choi, WildVis: Open Source Visualizer for Million-Scale Chat Logs in the Wild, 2024. URL: http://arxiv.org/abs/2409.03753. doi:10.48550/arXiv.2409.03753, arXiv:2409.03753 [cs].

[11] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019. URL: http://arxiv.org/abs/1908.10084, arXiv:1908.10084 [cs].

[12] K. R. Shahapure, C. Nicholas, Cluster Quality Analysis Using Silhouette Score, in: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020, pp. 747–748. URL: https://ieeexplore.ieee.org/document/9260048. doi:10.1109/DSAA49011.2020.00096.

[13] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27.

[14] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, J. Kaplan, Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022. URL: https://arxiv.org/abs/2204.05862v1.

[15] B. Wen, J. Yao, S. Feng, C. Xu, Y. Tsvetkov, B. Howe, L. L. Wang, Know Your Limits: A Survey of Abstention in Large Language Models, 2024. URL: http://arxiv.org/abs/2407.18418. doi:10.48550/arXiv.2407.18418, arXiv:2407.18418 [cs].

[16] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. URL: http://arxiv.org/abs/2203.02155. doi:10.48550/arXiv.2203.02155, arXiv:2203.02155 [cs].

[17] D. Misra, Mish: A Self Regularized Non-Monotonic Activation Function, 2020. URL: https://www.bmvc2020-conference.com/conference/papers/paper_0928.html.

[18] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework, 2019. URL: http://arxiv.org/abs/1907.10902, arXiv:1907.10902 [cs, stat].

[19] D. Vila-Suero, F. Aranda, Argilla - Open-source framework for data-centric NLP, 2023. URL: https://github.com/argilla-io/argilla.

[20] P. H. Martins, P. Fernandes, J. Alves, N. M. Guerreiro, R. Rei, D. M. Alves, J. Pombal, A. Farajian, M. Faysse, M. Klimaszewski, P. Colombo, B. Haddow, J. G. C. de Souza, A. Birch, A. F. T. Martins, EuroLLM: Multilingual Language Models for Europe, 2024. URL: https://arxiv.org/abs/2409.16235v1.

[21] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, 2017. URL: https://arxiv.org/abs/1711.05101v3.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, HuggingFace's Transformers: State-of-the-art Natural Language Processing, 2020. URL: http://arxiv.org/abs/1910.03771. doi:10.48550/arXiv.1910.03771, arXiv:1910.03771 [cs].

[23] W. Heeringa, C. Gooskens, V. J. van Heuven, Comparing Germanic, Romance and Slavic: Relationships among linguistic distances, Lingua 287 (2023) 103512. URL: https://www.sciencedirect.com/science/article/pii/S0024384123000360. doi:10.1016/j.lingua.2023.103512.