MaiChat: A Text-based Dialogue Corpus Rich In Conversational Features

Mai Hoang Dao, Catherine Lai, Peter Bell
The Centre for Speech Technology Research,
University of Edinburgh, United Kingdom
{Mai.Hoang.Dao, C.Lai, Peter.Bell}@ed.ac.uk

Abstract

We present a new English corpus of typed instant-messaging dialogues that includes detailed timing information. Messages are collected from interactions between pairs who know each other well; the corpus is rich in typed features that augment the purely lexical, including hesitations, self-corrections, expressive respellings, and other markers of spontaneous interaction. Messages are collected using a custom-built chat platform that logs not only message content but also keystroke dynamics, screen activity, and demographic metadata. Designed with a transparent and reproducible protocol, the corpus enables scalable data collection while ensuring privacy and consent. We intend that the rich collection of features collected will facilitate future research in areas such as cognitive modelling, human–computer interaction, and conversational AI.

Keywords: written conversation, keystroke dynamics, behavioral and contextual features

1. Introduction

Text-based communication, particularly in instant messaging, is rich in phenomena that extend beyond the literal content of messages. Users frequently employ hesitations, self-corrections, pauses, expressive respellings, repeated punctuation, and emoji use to convey affect, emphasis, or engagement. These behaviors reflect cognitive and social processes and are increasingly observed in digital dialogue (Wengelin, 2001; Molomer and Trausan-Matu, 2016; Medimorec et al., 2017).

While existing datasets in natural language processing and human-computer interaction have advanced research on text-based conversation, they often lack the integration of natural conversational flow with detailed behavioral and contextual information, such as keystroke timings, message edits, pauses, typing indicators, screen activity, and participant demographics. On the one hand, widely used text-based conversational datasets like NPS (Forsythand and Martell, 2007) or DailyDialog (Li et al., 2017) primarily capture message content. On the other hand, previous keystroke-based research has focused on, for example, authentication (Arsh et al., 2024; Sridhar et al., 2015), biometric analysis (Monrose and Rubin, 1997; Shadman et al., 2025), or other tasks (Borj and Bours, 2019; Ning et al., 2025; Buker et al., 2020; Lee et al., 2014), with no real human interactions and no publicly available data due to privacy concerns. For example, Borj and Bours (2019) explored whether keystroke patterns can indicate lying, but participants communicated via predefined scenarios rather than engaging in natural conversation. Similarly, (Ning et al., 2025) attempted to predict cognitive functioning in mood disorders through smartphone typing dynamics, but no public dataset is available, and the protocol is difficult to reproduce or scale due to its intensive time and effort requirements.

These limitations leave significant gaps for studying cognitive processes, emotion, engagement, social dynamics, and adaptive interfaces. To address them, we introduce a transparent and scalable data collection protocol for capturing anonymized, real-time spontaneous text-based conversations. Our corpus, ***MaiChat***, is enriched with behavioral and contextual features such as keystroke logs, message timing, typing indicators, screen activity, and demographic metadata, and captures interactional patterns including hesitations, expressive respellings, and non-lexical expressions such as emojis. Examples of typical conversational patterns observed in our corpus are shown in Figure 1.

In this paper, we present our data collection protocol and the conversational dialogue features that this protocol allows us to capture. We present quantitative and qualitative analyses of the data and compare this to other commonly used text-based dialogue corpora. We find that our data collection method successfully elicits features of spontaneous, conversational interaction, with features like typed laughter, emoji, and expressive respellings appearing much more prominently in MaiChat than other text chat datasets. We argue that these features align MaiChat more closely with corpora of (transcribed) spontaneous spoken dialogue, e.g., Switchboard (Godfrey and Holliman, 1993) and CallHome (Canavan et al., 1997). Nev-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1585–1594
11–16 May 2026.

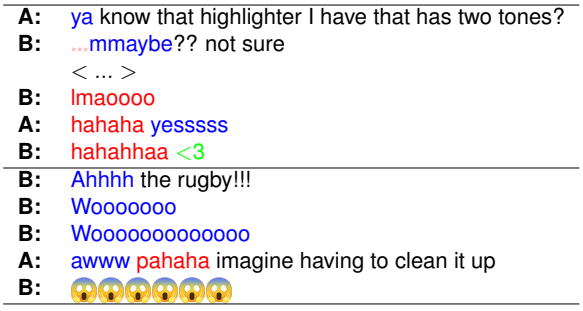

Figure 1: Examples of conversational patterns in our dataset. We add red for laughter, green for emoticons, blue for intentional misspellings, and pink for filled pauses.

ertheless, exploration of typing behaviours also shows how the text-based dialogue in MaiChat is distinct from spoken dialogue. We also find that LLMs alone fall short in generating the types of features observed in MaiChat. Moreover, while exposure to MaiChat data can make LLM generated chat more similar to MaiChat, we observe that feature diversity is still lower than in our human-human dialogues.

We hope that the additional information elicited by the MaiChat data collection process will advance understanding of the dynamics of text-based conversations across multiple domains. For example, we expect that detailed behaviour information will enhance investigations into cognitive and affective processes (Lee et al., 2014; Nahin et al., 2014; Poria et al., 2019; Pereira et al., 2024); human–computer and human–robot interaction, supporting the development of adaptive and socially aware systems (Iftikhar et al., 2023; Yang et al., 2021); conversational AI and chatbots, through modeling human-like interaction and turn-taking (Goodkind, 2021; Supriyanto et al., 2019); behavioral biometrics and security (Arsh et al., 2024); education and collaborative learning technologies (Farida et al., 2025; Sun et al., 2021; Talebinamvar and Zarrabi, 2022); and sociolinguistics and pragmatics, by facilitating analyses of variation, turn-taking, and digression in written social conversation (Anderson et al., 2007; Guydish and Fox Tree, 2022).

We publicly release the MaiChat dataset for research and educational purposes.[1]

2. Background

2.1. Written Conversational Features

Spoken discourse is typically characterized by spontaneity, interaction, and co-construction, whereas written discourse is more planned, structured, and permanent (Chafe, 1982; Chafe and Tannen, 1987; Halliday, 1989). Accordingly, the term "conversational features" traditionally evokes phenomena associated with spoken language. Yet, with the rise of online communication, many of these features now appear in written chat, where real-time, informal, and multi-threaded exchanges exhibit interactional patterns reminiscent of spoken interaction (Crystal, 2006; Garcia and Baker Jacobs, 1999; Meredith and Stokoe, 2014).

In general, online messaging is characterised by more dynamic/informal structure, as indicative of spoken interaction (Baron, 2010; Herring, 2013; Tagg, 2015). Moreover, digital platforms have created new conventions for interaction, including multimodal elements like emojis and GIFs, which function similarly to nonverbal cues in spoken conversation (Androutsopoulos, 2011). Several studies have observed conversational features in text-based interaction (König, 2019; Petitjean and Morel, 2017; Sampietro, 2021; Wengelin, 2001; Medimorec et al., 2017; Molomer and Trausan-Matu, 2016; Tarighat et al., 2024). However, there are still differences between modalities. For example, Molomer and Trausan-Matu (2016) finds that disfluency is more prevalent in spoken language compared to text-based chat. Similarly, studies have suggested that turn-taking in computer-mediated discourse differs substantially from that of oral conversation (Garcia and Baker Jacobs, 1999). These features provide rich signals about affect, cognition, and interaction style, which are often absent in traditional corpora.

To capture these aspects, we design the collection protocol of MaiChat to encourage conversational interaction, guided by the following observations how conversation differs from other forms of text chat/task-oriented dialogue:

- Informality: Participants engage without specific goals in a spontaneous and unplanned way, mirroring the informality of face-to-face conversations.
- Immediacy: The expectation of rapid responses leaves participants with less time to formulate their thoughts, encouraging active and continuous engagement.
- Non-linearity: Turn-taking structure can be unpredictable and non-linear, and multiple threads may develop simultaneously.

2.2. Related Text Chat Datasets

Many text chat corpora have been collected, spanning spontaneous chat rooms (Forsythand and Martell, 2007; Kummerfeld et al., 2019; Ringer et al., 2020), everyday conversations (Li et al., 2017), customer service interactions (Lowe et al., 2015; Axelbrooke, 2017), and task-specific chats

[1] Available at https://doi.org/10.7488/ds/8083.

(Zhang et al., 2018; Gopalakrishnan et al., 2019; Lewis et al., 2017). These datasets vary in conversationality, participants, and focus. For example, UDC (Lowe et al., 2015) is heavily task-oriented, DailyDialog (Li et al., 2017) lacks spontaneity, and NPS (Forsythand and Martell, 2007) differs in group chat structure. More recent corpora such as Topical-Chat (Gopalakrishnan et al., 2019) and PERSONA-CHAT (Zhang et al., 2018) involve strangers or topic/persona-specific dialogues. However, none of these corpora fully capture the temporal and behavioral dynamics of real-time interaction. Critically, these datasets do not record keystroke sequences, message timing, or pauses. With this in mind, we designed MaiChat to capture these sort of features to enable new analyses of cognitive and behavioral dynamics. We investigate these differences further in Section 4 and Section 5.

3. Dataset Design

We have collected conversations in the MaiChat corpus using a custom-built real-time chat platform.

3.1. Web-based Messenger Application

We developed our own messaging platform for the data collection. The platform's interface mirrors that of widely used messaging applications, such as Messenger or WhatsApp, to provide a user-friendly experience. The platform records:

- content and timing of each message sent.
- user keystroke timings during composition
- a screen (active) log to track when a user navigates away from the chat interface

We incorporated an "is-typing" indicator to improve participants' engagement and overall chatting experience (Iftikhar et al., 2023; Kalman and Rafaeli, 2011; Hancock and Dunham, 2001), filling a gap left by prior text-chat datasets where such a feature was not explicitly included.

3.2. Task

We aimed to create a relaxed setting in order to encourage open conversation from the start, whilst steering the topic away from personal details. To achieve this, we introduce brief, amusing videos covering everyday topics such as childhood incidents, pets, and other generally relatable themes. Participants were given cue questions to start conversation, following the video viewing, but it was emphasised they should feel free to digress onto any other topics of interest.

3.3. Participants

All participants are native English speakers , paired with partners from their social circles. This setup is designed to make the conversation as natural and enjoyable as possible. We record all interactions with the system anonymously and have ensured that the dataset can be publicly shared in a way that protects participants' privacy. MaiChat contains 42 conversations from 84 participants from diverse national backgrounds, including British, American, Canadian, and Indian, among others, totaling 40,016 tokens. The participant pool is 40.48% male, 53.57% female, and 5.95% other, with ages ranging from 18 to 61. Conversation pairs are primarily friends (71.4%) and romantic partners (22%), with the remainder being family members and colleagues. Examples of a snippet of our dataset are presented in Table 1.

Log	Content	Timing (hh:mm:ss.ms)
Chat log	Hi!	20:13:50.843
Typing log	H	20:13:48.789
	Hi	20:13:49.130
	Hi!	20:13:50.218
Screen log	isActive: TRUE	20:14:04.507
	isActive: FALSE	20:15:57.210

Table 1: Snippets of our dataset. Samples are simplified for display purposes.

4. Analysis of Conversational Content

This section analyses the message-level content of MaiChat to highlight its unique conversational characteristics. By comparing with other chat corpora and LLM-generated dialogues, we aim to show how MaiChat better captures the spontaneity and affective richness of natural human communication.

4.1. Comparison to Other Datasets

We compare several text chat datasets, which we believe represent the most "conversation-like" chat data collected to date: NPS, Topical-Chat, PERSONA-CHAT, DailyDialog, UDC, Customer Support on Twitter, Deal or No Deal, IRC, TwitchChat. For speech data, we use a combination of Switchboard (SWB) and CallHome corpora, which are commonly used together based on similarities in their data collection protocols. These corpora are both spontaneous nature, and come with the careful transcription of laughter, disfluencies and filled pauses. However, we note that speakers were strangers in SWB, which may effect their interaction style. Given the significant size variation

Statistics	# tokens	# laughter	# OL	# AL	# EM	# TE	# FP	# IM	# IC
MaiChat	40016	412	253 [83]	159 [39]	274 [114]	84 [37]	91 [26]	326	135
NPS	40153	732	42 [25]	690 [5]	10 [3]	93 [24]	26 [12]	128	19
Topical	40073	172	92 [8]	80 [3]	0	18 [5]	0	7	0
PERSONA-CHAT	39985	115	43 [3]	72 [1]	0	0	7 [3]	8	0
DailyDialog	40112	0	0	0	0	0	6 [2]	0	0
Ubuntu Dialogue Corpus	40106	121	113 [12]	8 [1]	0	197 [27]	48 [7]	29	33
Customer Support on Twitter	40829	12	7 [1]	5 [1]	349 [82]	25 [10]	0	3	0
Deal or No Deal	39022	18	7 [1]	11 [1]	0	0	8 [2]	16	25
IRC	40172	618	43 [23]	575 [24]	0	67 [15]	18 [3]	132	17
TwitchChat	40190	189	135 [9]	54 [7]	326 [37]	61 [15]	3 [1]	23	19
Zero-shot GPT	40055	153	96 [2]	57[1]	352 [84]	0	0	19	18
Few-shot GPT	40211	314	242 [12]	72 [13]	393 [64]	5 [2]	31 [12]	72	149
Switchboard	40024	237	N/A	N/A	N/A	N/A	1832 [7]	N/A	N/A
CallHome	40015	259	N/A	N/A	N/A	N/A	158 [7]	N/A	N/A

Table 2: Conversational phenomena statistics of our dataset. OL: onomatopoeic laughter, AL: other laughter, EM: emojis, TE: typed emoticons, FP: filled pauses, IM: intentional misspellings, IC: intentional capitalisation, Numbers in square brackets indicate the number of distinct examples seen from each category.

across corpora, we ensure comparability by extracting 40K tokens random subsets from each corpus, matching the total token count in our dataset. In the following, we report the averages from 5 randomly chosen subsets per corpus.

Manual annotations of various conversational features in MaiChat and other datasets were performed by a single annotator. Our goal is not to provide detailed annotations in itself, but rather to provide an overview of systematic differences across corpora. With that in mind, a single annotator was deemed sufficient. However, we also performed spot checks which revealed no systematic errors.

Table 2 provides statistics on the occurrence of conversational phenomena. MaiChat exhibits a substantially higher occurrence of laughter than the other corpora, both onomatopoeic (e.g., "haha" "hehe") and other forms (e.g., "lol" "lmao"). NPS and IRC do surpass our dataset in the total number of other laughter occurrences, but these are of a limited variety.

Our dataset is also very high in emoji usage, with 274 instances, similar to other public chat or chat-room datasets such as Twitter, and Twitchat. MaiChat includes a wide range of emojis: 114 distinct instances of emojis and 37 distinct instances of typed emoticons–much more diverse than most human-generated chat datasets. While filled pauses, misspellings and non-standard capitalisation are almost non-existent in other text-based chat corpora, they are much more frequent in MaiChat. This further emphasises its richness in conversational features. In the following, we further discuss the use of the conversational phenomena in MaiChat, and how this differs from LLM generated chat.

4.2. Laughter, Emojis & Typed Emoticons

Figure 1 presents examples of the laughter and emoji patterns observed in our dataset. Unlike other chat corpora, our dataset captures a wide array of laughter expressions (e.g., "lmaoooo" "hahaha") and emojis, reflecting the diverse ways people express emotions in digital conversations. Similarly, the deployment of emojis, such as the series of shocked faces "😱😱😱", convey complex emotions and reactions that might be challenging to express through words alone.

4.3. Expressive Respelling

We find numerous instances of "respelling" in our data, i.e., where participants enrich the text-based interaction by deviating from formal written text. For example, in the snippet shown in Figure 1, participant B uses "..." followed by the intentional misspelling "mmaybe" to indicate uncertainty.

Intentional misspellings in MaiChat (cf. Figure 1) mirror prosodic variation in spoken language. For example, "mmaybe" represents the speech elongation that often occurs when someone is unsure of their response. Duplication in "yesssss" expresses enthusiastic confirmation or joy, similar to how vocal emphasis is used to convey strong emotions. Similarly, capitalisation, e.g. "SO GOOD", like vocal emphasis, cues heightened emotional states or shifts in the expressive intensity, underscoring the significance of the content.

4.4. Comparison to LLM Generated Chats

We also examine the ability of GPT-4 (OpenAI et al., 2024) to generate realistic human-like writ-

Mai Chat	Nah nah you'd love it 😂 😂😂😂😂it acc looksss superrrr delicious 🤪u wish the ending was 🤯
GPT-4	I'm now officially a matcha master chef. 😌🍵 I finally beat that level in Super Mario Odyssey I was stuck on for weeks! 🎮🌟 Your favorite chocolate cake coming right up. 🎂

Figure 2: Examples of emojis used in MaiChat and GPT-4.

ten conversations. This evaluation serves two purposes: it demonstrates how conversational data like MaiChat can benchmark dialogue naturalness, and it highlights current gaps between model-generated and human-written interactions. In doing so, we connect corpus design with downstream applications in dialogue modeling and human–AI communication research. Two settings are used: (i) zero-shot, where the model generates dialogues without access to real data, and (ii) few-shot, where it continues real MaiChat conversations after seeing their first 10 lines. The model is prompted with the same task instructions as human participants, simulating different relationship types. Overall, GPT-4 seems to struggle to sustain long, natural exchanges comparable to the 15-minute human dialogues. To ensure size comparability, we generated 123 and 61 dialogues for the zero-shot and few-shot settings, each totaling around 40K tokens.

MaiChat demonstrates substantially richer expressive behaviour than the Zero-shot GPT-4 dataset, with more laughter expressions both in frequency (412 vs. 153 instances) and variety (122 vs. 2 types), indicating greater humour and spontaneity. Emoji use is also more diverse and meaningful: 30.17% of GPT-4 emojis are merely illustrative (e.g., 🎂, 🌟, 🎮), compared to 3.6% in MaiChat, where emojis often express or intensify emotions and even substitute for words (e.g., "the ending was 🤯"–i.e., "the ending was mind-blowing"– giving a direct representation of feelings and providing insight into the speaker's state beyond the textual content). GPT-4 outputs also lack intentional misspellings and filled pauses, distinguishing it from our human-human chat.

Quantitatively, the few-shot setting increases the frequency of several features compared to zero-shot, including onomatopoeic laughter, intentional misspellings, and capitalization. Emojis are also more naturally integrated, with only 17.2% being illustrative, and filled pauses and typed emoticons–virtually absent in zero-shot outputs–do appear.

Despite these improvements, feature diversity remains far lower than in MaiChat (e.g., 12 vs. 83 distinct laughter instances), indicating that few-shot GPT tends to mimic rather than generalise from examples. Overuse of forms such as "haha" (164+ instances) and excessive emojis or intentional misspellings further suggests reliance on surface cues from the prompt rather than a deeper replication of natural discourse patterns.

Statistics	**Counts**
Total messages	5297
Avg messages per conversation	125
Median message duration	19.37 s
Total turns	2856
Median turn duration	31.57 s
Median turn duration after pause partner	33.27 s
Median turn duration after fast partner	25.45 s
Shortest turn duration	0.05 s
Longest turn duration	61.12 s
Average turns per conversation	68
Total keystrokes	89728
Median keystroke duration	210 ms
Avg keystrokes per conversation	2136
IKI distribution	220 ms; 210 ms;1.1;3.3
Typing acceleration/deceleration	$\pm$ 50 ms^2

Table 3: General statistics of the MaiChat's composition phase. For inter-keystroke intervals (IKI), mean, median, skewness, and kurtosis are reported in this order.

5. Analysis of Composition Phase

5.1. General Statistics

The keystroke-level log provides a detailed view of participants' interaction dynamics (Table 3). We define a turn as the interval starting from a participant's first message in a sequence of one or more consecutive messages and ending at the time of the partner's first message in their subsequent sequence. Even a single message constitutes a turn under this definition. This approach captures the temporal structure of conversation, producing very short turns for rapid interjections and longer turns when participants are inactive or the partner delays responses. On average, each conversation contains 125 messages and 68 turns, with turn durations ranging from 0.05 s–i.e. very short turns contain single message such as "ok", "ye"–to 61.12 s, reflecting rapid exchanges alongside occasional longer, more deliberate turns. Median keystroke duration is 210 ms, and the inter-keystroke interval (IKI) distribution exhibits a slight positive skew (1.1) with a moderately heavy tail (kurtosis 3.3), indicating generally consistent typing speeds punctuated by occasional longer pauses.

We observe that turn durations vary with the partner's recent activity. Turns following a pause between messages in the partner's turn (> 5 times average interval) are slightly longer than the median (33.27 s vs. 31.57 s), while turns after rapid consecutive messages (no pause) are shorter (25.45 s vs. 31.57 s). These patterns, though not causal, show how MaiChat's temporal logs capture interac-

Statistics	Counts
Messages w/ TB, no AS change	3182
Messages w/ TB, AS change	84
Typing behaviours	4277
Delays	1521
Deletions	2355
Insertions	401
Messages w/ three TB types	168
Partial-word deletions	789
Full-word deletions	1,312
Undo patterns	254
Laughter w/ TB	51

Table 4: Statistics of typing behaviours. AS denotes "Active Status". TB denotes "Typing Behaviour".

Type	Examples
	They're friendly and loveable
delay	They're friendly and loveable
deletion	friendly and loveable
insertion	Dogs are friendly and loveable
insertion	Dogs are loyal and friendly and loveable

Table 5: Example of a sequential typing behaviours from MaiChat.

tional and partner-influenced dynamics in real-time conversation.

This fine-grained timing data enables the study of typing rhythm, micro-level typing behaviour patterns, and conversational pacing that are often invisible in standard text-based dialogue corpora.

5.2. Typing Behaviours

The keystroke logging in MaiChat allows us to analyse typing behaviours, which are in some ways akin to the concept of disfluencies in spoken language (Medimorec et al., 2017; Wengelin, 2001). While some instances correspond to simple typos, we also observe more substantial hesitations, retractions, and corrections, reminiscent of the processes found in spoken conversation. However, in spoken dialogue we can only observe the what was actually said, but with typed chat, we can also observed what was written but then deleted or inserted. This allows for a different view of message construction and turn-taking.

Following Shriberg (1994)'s categorisation of speech disfluencies, we define typing behaviours as the temporal and revision-related phenomena observed during typing: (i) messages exhibiting at least one significant *delay* between keystrokes (> 5 times the average interval); (ii) messages involving the *deletion* of at least one character during composition; and (iii) messages with *insertion* of at least one character into an existing typed sequence.

Composition behaviours were observed in 60% of messages (Table 4), indicating a highly spontaneous and uncertain typing flow. Among these, deletions were the most frequent, occurring in 44% of messages, though all three types could co-occur within the same message. Table 5 illustrates the different types of typing behaviours within a single message. The pause reflects hesitation during typing, potentially signalling momentary uncertainty about what to write. Then the deletion of "they're" indicates rephrasing and the insertion represents the addition of information to improve clarity.

Out of 5,297 messages, 3,182 exhibiting typing behaviours were typed without any change in active status, indicating that the participant was fully focused. Within these messages, 1,521 delays, 2,355 deletions, and 401 insertions were observed. Partial-word deletions account for 33.5% of deletions, whereas full rewrites are less frequent (10.7%). Interestingly, 12.3% of laughter-containing messages also include edits or delays, suggesting that even playful messages are sometimes refined. These keystroke-level features provide insights into conversational style and micro-level behaviours that cannot be captured by standard message timestamps.

6. Discussion and Conclusion

MaiChat contains a rich set of features that reflect natural interactional and expressive dynamics, including misspellings and capitalization reminiscent of prosodic cues, filled pauses, self-corrections, and onomatopoeic expressions that signal hesitation and emphasis. Quantitative analyses show that the frequency of these markers aligns with spoken dialogue corpora such as Switchboard (Godfrey and Holliman, 1993) and CallHome (Canavan et al., 1997), highlighting MaiChat's capacity to capture conversationality, interactivity, and spontaneity. Beyond textual content, its detailed keystroke-level logs, message timing, and contextual metadata enable research across multiple domains, including cognitive science, human–robot interaction, sociolinguistics, and conversational AI.

While recent large language models generate coherent text, their outputs often lack the nuanced, spontaneous, and expressive patterns observed in human dialogue (Mayor et al., 2025; Jawale et al., 2024). Our few-shot experiments demonstrate that exposure to MaiChat can guide LLMs toward more varied, cognitively plausible, and human-like dialogue in timing, interactivity, and expressiveness, underscoring the value of rich behavioral and contextual information.

In summary, MaiChat provides a novel corpus of written dialogue that captures a broader spectrum of conversational markers than existing text-based

corpora and LLM-generated dialogues. Its combination of textual, behavioral, and contextual features offers a transparent and openly accessible resource for advancing research in linguistics, cognitive science, human–computer interaction, and conversational AI, providing a unique tool for understanding human communication in interactive written contexts.

7. Limitations and Potential Risks

7.1. Limitations

Although MaiChat offers rich conversational features and our experiments demonstrate its alignment with spoken speech, particularly SWB, compared to other chat datasets, there are some limitations to our work. First, MaiChat is smaller than other text-chat datasets. However, it was collected with careful attention to ethical considerations, using paid participants on a custom messaging platform, making the data collection process more resource-intensive but yielding a richer source of conversational dynamics.

Second, several important features, such as timing, emojis, and typing patterns, remain unexplored and could provide further insights. These elements may be better captured by more advanced clustering techniques or specialized models.

Moreover, all conversational features were annotated by a single annotator. In practice, the tagging guidelines proved straightforward and internally consistent, and spot-checks during post-processing revealed no systematic errors; we therefore do not expect additional annotators to materially change the findings. Nevertheless, future work could verify these annotations through an independent second pass or crowdsourced adjudication to quantify reliability and further reduce any residual noise.

In conclusion, while our results are informative, further exploration using more advanced methods is necessary to gain a deeper understanding of the unique characteristics of MaiChat and its relation to other datasets

7.2. Potential Risks

There are no significant risks associated with this study, as participants engage in text-based conversations on a private messaging platform in a controlled environment. The study protocol was subject to an independent ethics review, ensuring that all ethical considerations regarding participant privacy and data security were met.

There remains a low potential risk that participants may be identified based on the information they share during the conversation. To mitigate this, participants are informed of this risk in advance and advised not to disclose any personally identifiable information during their discussions. Additionally, the platform does not collect or store sensitive personal data beyond what is necessary for the study. Participants are free to discuss any topic they feel comfortable with, and if they feel uneasy at any point, they may stop the conversation and request the data to be removed. Given these precautions, we believe the risks are minimal and well-managed.

In addition, training models on the MaiChat corpus–which consists of genuine human conversations–introduces a subtle societal impact: the model may sound so natural that telling its messages apart from genuine chat becomes harder, potentially complicating fraud or impersonation checks. Future work should pair this training data with simple safeguards (e.g., provenance tags or light watermarking) and assess models not only for accuracy but also for their impact on the detectability of machine-generated content.

8. Bibliographical References

Jeffrey F Anderson, Fred K Beard, and Joseph B Walther. 2007. Turn-taking and the local management of conversation in a highly simultaneous computer-mediated communication system.

Jannis Androutsopoulos. 2011. Language change and digital media: A review of conceptions and evidence. *Standard languages and language standards in a changing Europe*, 1:145–159.

Aditya Arsh, Nirmalya Kar, Smita Das, and Subhrajyoti Deb. 2024. Multiple approaches towards authentication using keystroke dynamics. *Procedia Computer Science*, 235:2609–2618.

Stuart Axelbrooke. 2017. Customer support on twitter.

Naomi S Baron. 2010. *Always on: Language in an online and mobile world*. Oxford University Press.

Parisa Rezaee Borj and Patrick Bours. 2019. Detecting liars in chats using keystroke dynamics. In *Proceedings of the 2019 3rd international conference on biometric engineering and applications*, pages 1–6.

Abeer AN Buker, Giorgio Roffo, and Alessandro Vinciarelli. 2020. Type like a man! inferring gender from keystroke dynamics in live-chats. *IEEE Intelligent Systems*, 34(6):53–59.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME American English Speech LDC97S42. Linguistic Data Consortium.

W Chafe. 1982. Integration and involvement in speaking, writing, and oral literature. *Spoken and written language: Exploring orality and literacy/Ablex*, 3554.

Wallace Chafe and Deborah Tannen. 1987. The relation between written and spoken language. *Annual review of anthropology*, 16:383–407.

David Crystal. 2006. *Language and the Internet*.

Jaha Farida, Kartit Ali, and Fertat Mohamed. 2025. Leveraging keystroke dynamics to detect identity fraud and ai-driven cheating in online education. In *International Symposium on Generative AI and Education*, pages 487–498. Springer.

Eric N. Forsythand and Craig H. Martell. 2007. Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing (ICSC 2007)*, pages 19–26.

Angela Cora Garcia and Jennifer Baker Jacobs. 1999. The eyes of the beholder: Understanding the turn-taking system in quasi-synchronous computer-mediated communication. *Research on language and social interaction*, 32(4):337–367.

John Godfrey and Edward Holliman. 1993. Switchboard-1 Release 2 LDC97S62. *Linguistic Data Consortium*.

Adam Goodkind. 2021. *Using keystrokes to predict social dynamics in dialogue*. Ph.D. thesis, PhD Dissertation, Northwestern University, 2023, https://adamgoodkind. com ....

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Andrew J Guydish and Jean E Fox Tree. 2022. Reciprocity in instant messaging conversations. *Language and speech*, 65(2):404–417.

Michael Alexander Kirkwood Halliday. 1989. Spoken and written language.

Jeffrey T Hancock and Philip J Dunham. 2001. Impression formation in computer-mediated communication revisited: An analysis of the breadth and intensity of impressions. *Communication research*, 28(3):325–347.

Susan C Herring. 2013. Discourse in web 2.0: Familiar, reconfigured, and emergent. *Discourse*, 2(0):1–26.

Zainab Iftikhar, Yumeng Ma, and Jeff Huang. 2023. "together but not together": Evaluating typing indicators for interaction-rich communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Toshish Jawale, Chaitanya Animesh, Sekhar Vallath, Kartik Talamadupula, and Larry Heck. 2024. Are human conversations special? a large language model perspective. *arXiv preprint arXiv:2403.05045*.

Yoram M Kalman and Sheizaf Rafaeli. 2011. Online pauses and silence: Chronemic expectancy violations in written computer-mediated communication. *Communication Research*, 38(1):54–69.

Katharina König. 2019. Stance taking with 'laugh'particles and emojis–sequential and functional patterns of 'laughter'in a corpus of german whatsapp chats. *Journal of Pragmatics*, 142:156–170.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Po-Ming Lee, Wei-Hsuan Tsui, and Tzu-Chien Hsiao. 2014. The influence of emotion on keyboard typing: an experimental study using visual stimuli. *Biomedical engineering online*, 13(1):81.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Eric Mayor, Lucas M Bietti, and Adrian Bangerter. 2025. Can large language models simulate spoken human conversations? *Cognitive Science*, 49(9):e70106.

Srdan Medimorec, Torin P Young, and Evan F Risko. 2017. Disfluency effects on lexical selection. *Cognition*.

Joanne Meredith and Elizabeth Stokoe. 2014. Repair: Comparing facebook 'chat'with spoken interaction. *Discourse & communication*, 8(2):181–207.

Sibel Denisleam Molomer and Stefan Trausan-Matu. 2016. Analysis of disfluency in audio and chat transcripts. In *Proceedings of the 20th International Conference on System Theory, Control and Computing (ICSTCC)*.

Fabian Monrose and Aviel Rubin. 1997. Authentication via keystroke dynamics. In *Proceedings of the 4th ACM Conference on Computer and Communications Security*, pages 48–56.

AFM Nazmul Haque Nahin, Jawad Mohammad Alam, Hasan Mahmud, and Kamrul Hasan. 2014. Identifying emotion by keystroke dynamics and text pattern analysis. *Behaviour & Information Technology*, 33(9):987–996.

Emma Ning, Ryne Estabrook, Theja Tulabandhula, John Zulueta, Mindy K Ross, Sarah Kabir, Faraz Hussain, Scott A Langenecker, Olusola Ajilore, Alex Leow, et al. 2025. Predicting cognitive functioning in mood disorders through smartphone typing dynamics. *Journal of psychopathology and clinical science*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and Lama Ahmad et al. 2024. Gpt-4 technical report.

Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. 2024. Deep emotion recognition in textual conversations: A survey. *Artificial Intelligence Review*, 58(1):10.

Cécile Petitjean and Etienne Morel. 2017. "hahaha": Laughter as a resource to manage whatsapp conversations. *Journal of Pragmatics*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

Charles Ringer, Mihalis Nicolaou, and James Walker. 2020. Twitchchat: A dataset for exploring livestream chat. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 16, pages 259–265.

Agnese Sampietro. 2021. Emojis and the performance of humour in everyday electronically-mediated conversation: A corpus study of whatsapp chats. *Internet Pragmatics*, 4(1):87–110.

Rashik Shadman, Ahmed Anu Wahab, Michael Manno, Matthew Lukaszewski, Daqing Hou, and Faraz Hussain. 2025. Keystroke dynamics: Concepts, techniques, and applications. *ACM Computing Surveys*, 57(11):1–35.

Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California.

Mahalaxmi Sridhar, Siddhesh Vaidya, and Piyush Yawalkar. 2015. Intrusion detection using keystroke dynamics & fuzzy logic membership functions. In *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, pages 1–10. IEEE.

Bo Sun, Yong Wu, Kaijie Zhao, Jun He, Lejun Yu, Huanqing Yan, and Ao Luo. 2021. Student class behavior dataset: a video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. *Neural Computing and Applications*, 33(14):8335–8354.

Supriyanto, Adhi Prahara, and Tri Susanto Saputro. 2019. Keystroke- level model to evaluate chatbot interface for reservation system. In *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pages 241–246.

Caroline Tagg. 2015. *Exploring digital communication: Language in action*. Routledge.

Mobina Talebinamvar and Forooq Zarrabi. 2022. Clustering students' writing behaviors using keystroke logging: a learning analytic approach in efl writing. *Language Testing in Asia*, 12(1):6.

Aida Tarighat, Patrick Sturt, and Martin Corley. 2024. Perspectives on language model and human handling of written disfluency and nonliteral meaning. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue*.

Asa Wengelin. 2001. Disfluencies in writing-are they like in speaking? In *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*.

Fangkai Yang, Yuan Gao, Ruiyang Ma, Sahba Zojaji, Ginevra Castellano, and Christopher Peters. 2021. A dataset of human and robot approach behaviors into small free-standing conversational groups. *PloS one*, 16(2):e0247364.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.