Comparing Large Language Models for Automated Subject Line Generation in e-Mental Health: A Performance Study

Philipp Steigerwald[a] and Jens Albrecht[b]
Technische Hochschule Nürnberg Georg Simon Ohm, Nürnberg, Germany
{*philipp.steigerwald, jens.albrecht*}*@th-nuernberg.de*

Keywords: Large Language Models, e-Mental Health, Psychosocial Online Counselling, Subject Line Generation, Inter-Rater Reliability.

Abstract: Large Language Models (LLMs) have the potential to enhance e-mental health and psychosocial e-mail counselling by automating tasks such as generating concise and relevant subject lines for client communications. However, concerns regarding accuracy, reliability, data privacy and resource efficiency persist. This study investigates the performance of several LLMs in generating subject lines for e-mail threads, yielding a total of 253 generated subjects. Each subject line was assessed by six raters, including five counselling professionals and one AI system, using a three-category quality scale (Good, Fair, Poor). The results show that LLMs can generally produce concise subject lines considered helpful by experts. While GPT-4o and GPT-3.5 Turbo outperformed other models, their use is restricted in mental health settings due to data protection concerns, making the evaluation of open-source models crucial. Among open-source models, SauerkrautLM LLama 3 70b (4-bit) and SauerkrautLM Mixtral 8x7b (both 8-bit and 4-bit versions) delivered promising results with potential for further development. In contrast, models with lower parameter counts produced predominantly poor outputs.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

LLMs have emerged as powerful tools for text summarisation and content generation (Wu et al., 2024; Zhang et al., 2024b). In the context of e-mental health and psychosocial e-mail counselling, these models could provide valuable support by automatically suggesting alternative subject lines alongside the original ones, offering counsellors additional context for client communications. However, deploying LLMs in sensitive domains like e-mental health requires careful consideration. Inaccurate or biased content generation could lead to misunderstandings in counselling contexts (Chung et al., 2023; Guo et al., 2024; Li et al., 2024), necessitating thorough evaluation before implementation (Lawrence et al., 2024; Xu et al., 2024). This study investigates which LLMs can generate concise and relevant subject lines for psychosocial counselling communications and evaluates their potential for practical implementation. The analysis compares proprietary against open-source models, two quantisation levels and standard versus German language-tuned versions. From this comprehensive evaluation, the investigation addresses three fundamental research questions:

1. Are LLMs capable of effectively condensing social counselling e-mails into concise, meaningful one-line summaries, in other words subject lines?
2. Does fine-tuning models on the target language (German) result in improved output quality?
3. How do model size and quantisation impact the quality of generated subject lines?

This research evaluates LLMs for psychosocial e-mail counselling, a subset of e-mental health. Due to data privacy regulations in e-mental health settings, particular emphasis is placed on identifying secure open-source alternatives to proprietary models. Through comprehensive testing across different configurations the study aims to help institutions select solutions that optimally balance performance, language support and resource requirements.

[a] https://orcid.org/0009-0002-5564-4279
[b] https://orcid.org/0000-0003-4070-1787

Steigerwald, P. and Albrecht, J.
Comparing Large Language Models for Automated Subject Line Generation in e-Mental Health: A Performance Study.
DOI: 10.5220/0013294100003938
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 11th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2025)*, pages 70-77
ISBN: 978-989-758-743-6; ISSN: 2184-4984
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

2 RELATED WORK

LLMs are continually improving and expanding into more fields. In e-mental health, LLMs could assist therapists and counsellors by helping them quickly understand the core issues of clients from their initial inquiries. This section provides an overview of the literature on text summarisation, the application of LLMs in e-mental health contexts and evaluation methods for and with LLMs in sensitive domains.

2.1 Text Summarisation

Summarisation techniques are broadly classified into extractive and abstractive methods, each with distinct strengths and limitations (Zhang et al., 2024a). Recent research has introduced a hierarchical approach for summarising long texts that exceed the maximum input length of language models (Yin et al., 2024). This method involves topic-specific segmentation, condensation of segments and abstractive final summarisation. In the context of e-mail summarisation, previous work has demonstrated the feasibility of using AI to generate short, concise subject lines through a two-step approach (Zhang and Tetreault, 2019). This method first extracts key sentences from the e-mail text before rewriting them into concise subject lines. Further applications of summarisation in a healthcare setting have been explored, where LLMs are used to create short summaries of scientific abstracts for supporting clinical decision-making (Kocbek et al., 2022).

2.2 LLMs in Mental Health

Advancements in LLMs have expanded AI's potential in e-mental health. Systems such as ChatCounselor (Liu et al., 2023), Psy-LLM (Lai et al., 2024), MentalBlend (Gu and Zhu, 2024) and a ChatGPT-based approach (Vowels et al., 2024) aim to conduct realistic counselling sessions and simulate specific therapeutic techniques. However, these systems are not yet capable of autonomously replacing human counsellors or therapists (Chiu et al., 2024; Koutsouleris et al., 2022).

Additionally, systems like Reply+ (Fu et al., 2023) and CARE (Hsu et al., 2023) aim to support counsellors by providing suggestions or assisting in decision-making processes.

2.3 Evaluation of LLM Outputs

Evaluating LLMs in sensitive contexts like e-mental health requires robust assessment frameworks. While traditional metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) exist, they often fail to capture the nuances of mental health-related language. Human evaluation thus remains the gold standard for assessing LLM outputs in sensitive domains (Tam et al., 2024). Recent studies in healthcare have demonstrated successful evaluation approaches. A study of LLM diagnostic capabilities achieved high inter-rater reliability among medical professionals (Khan and O'Sullivan, 2024), while another implementation used Krippendorff's Alpha to assess AI-generated counselling responses (Rudolph et al., 2024). Research has also explored using LLMs themselves as evaluators, showing promising results in both educational assessment (Hackl et al., 2023) and argument quality analysis (Mirzakhmedova et al., 2024). The present study builds upon these evaluation approaches by employing both human experts and an AI system to rate the quality of subject lines generated by LLMs in the context of e-mental health.

3 METHODOLOGY

This study evaluates the generation of subject lines for psychosocial counselling e-mails across 11 different LLMs (see Table 1). Using 23 distinct e-mail threads, each model generated subject lines, yielding a total of 253 outputs.

Table 1: Overview of models used in the study. The FP column refers to full-precision (non-quantized) models, while Q4 and Q8 refer to models quantized with 4-bit and 8-bit precision, respectively.

Model Name	FP	Q4	Q8
GPT 3.5 Turbo	x		
GPT-4o	x		
Meta Llama 3.1 8b		x	x
SauerkrautLM Llama 3.1 8b		x	x
SauerkrautLM Llama 3 70b		x	
SauerkrautLM Mixtral 8x7b		x	x
Mistral Mixtral 8x7b		x	x

To ensure data privacy, counselling practitioners crafted e-mail content that simulates realistic counselling scenarios. Six raters — including five professionals in psychosocial online counselling and OpenAI's o1-preview model — assessed the generated subject lines using a three-category scale: Good, Fair and Poor. This process yielded a total of 1,518 ratings. Figure 1 shows an exemplary e-mail and different generated subject lines.

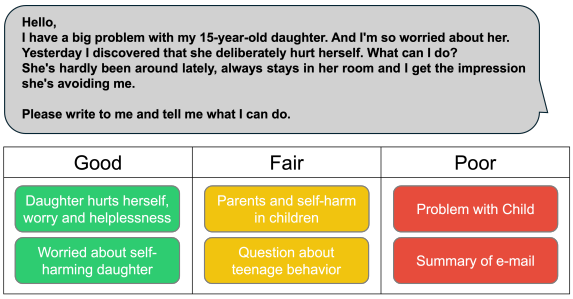

Figure 1: Example rating of generated subject lines for a counselling e-mail concerning self-harm (translated from German). Six generated subject lines are shown categorised into Good (capturing the urgent nature and specific issue), Fair (mentioning general topic) and Poor (overly generic) ratings. The original e-mail content is shown above.

3.1 LLM Selection

The selection of the models presented in Table 1 was guided by several key criteria to ensure a comprehensive evaluation across different categories of LLMs. Firstly, OpenAI's GPT series, specifically GPT-3.5 Turbo and GPT-4o, were included as they represent the leading proprietary models in the field (Shahriar et al., 2024; Rao et al., 2024), serving as benchmarks for high-performance language generation. In addition to OpenAI's models, Meta's Llama 3.1 8b (Dubey et al., 2024) and Mistral Mixtral 8x7b (Jiang et al., 2024) were selected as prominent open-source alternatives. VAGO's SauerkrautLM models (hereafter referred to as SKLM) provide German-tuned versions of these models, with SKLM Llama 3.1 8b and SKLM Mixtral 8x7b representing their fine-tuned counterparts. Both 4-bit and 8-bit quantized versions to assess the effects of quantisation on model performance were utilised. This approach allows to examine potential trade-offs between computational efficiency and output quality, providing a more nuanced view of how quantisation impacts model effectiveness. Finally, SKLM Llama 3 70b in a 4-bit quantisation was included to introduce a model with a larger parameter count. This addition allows to evaluate whether a mid-range model size offers notable performance advantages, providing a broader understanding of how model complexity affects outcomes in this task. Together, this selection enables a comprehensive analysis across a spectrum of proprietary, open-source and language-specific models with varying sizes and quantisations.

3.2 Subject Generation

The prompt followed a structured format defining the LLM's role in counselling, establishing context and setting, specifying the subject line generation task, detailing required input/output formats and noting formalities to avoid. After these specifications, the actual e-mail data was inserted according to the defined format. The prompt concluded with role/task reminders to maintain focus. The full prompt is presented in the following:

Generated Subject Prompt

You are a specialised assistant for psychosocial online counselling.

Clients often approach counselling services with vague subject lines like "Help" or "Problem." Your role is to assist the counsellor by generating a precise and individual subject line for the client's first e-mail. This helps the counsellor quickly grasp the main content of the request and respond efficiently, especially when managing multiple parallel cases.

Carefully read the client's first e-mail and generate a concise subject line that clearly and understandably summarises the core issue of the request. The subject should be a maximum of 6 words and should not contain unnecessary formalities, enabling the counsellor to immediately gain a clear understanding of the issue.

The input consists of a complete e-mail thread in chronological order. The e-mail is formatted as: {Role} wrote on {Date} at {Time}: 'e-mail Content' ###.

The desired output is a JSON object containing one field: { "Subject": "Generated concise subject line" }.

The subject line should concisely summarise the core content of the client's first message and avoid unnecessary formalities. Do not use quotation marks or 'Subject:' in the generated subject.

Following the formatted e-mail history is presented:
{{complete_e-mail_history}}
End of e-mail history.

Remember, you are a specialised assistant for psychosocial online counselling. Your task is to create concise and relevant subject lines that help the counsellor to quickly understand the client's issue.

Remember, your task is to read the client's first e-mail in the thread and generate a short, concise subject line that accurately reflects the core content of the request.

To ensure outputs conformed to a predefined JSON schema, each model's specific structured output capabilities were utilized accordingly. The required format was defined as:

Generated Subject Output Schema (JSON)

{"Subject": "Concise subject summarising core issue"}

3.3 Rater Line-up

The evaluation of the generated subject lines involved six raters consisting of five human experts in psychosocial online counselling and OpenAI's o1-preview, chosen for its strong reasoning capabilities (Temsah et al., 2024). Each rater independently assessed all 253 subject lines, categorising them into one of three quality categories — Good, Fair or Poor.

To ensure unbiased and consistent evaluations, both human-raters and the AI-rater received the client's initial e-mail along with all 11 generated subject lines for each e-mail, presented in random order and without model identifiers. This setup allowed raters to assess each subject line individually on an absolute scale while also considering them in relation to the other subject lines within the same e-mail context. The evaluators were given the following guidelines on which to base their evaluations:

> Subject lines should be concise and individually tailored to the initial message of the person seeking advice.
>
> Each subject line must summarise the main content clearly and understandably in a maximum of 6 words, avoiding unnecessary formalities.
>
> The evaluation focuses on how precisely and directly a subject line captures the core content of the initial message.
>
> High-quality subjects should enable counsellors to quickly grasp the central concern.

To ensure that the AI-rater evaluated each subject line correctly and did not inadvertently modify or paraphrase them, each subject line was associated with a unique hash value. The prompt instructed the AI-rater to rate each subject line and return a valid JSON object containing the hash and the assigned category.

3.4 Rating Agreement Analysis

To evaluate the consistency of the ratings, two complementary metrics were used: Spearman correlation and Krippendorff's Alpha. Spearman correlation is utilized to analyse the monotonic relationship between pairs of raters' rankings, while Krippendorff's Alpha assesses the overall reliability across all raters. Spearman correlation (Equation 1) is a non-parametric measure that evaluates whether two raters tend to rank subjects in a similar order, regardless of the absolute values assigned. In the context of this study, each rater's categorical ratings — Good, Fair and Poor — are assigned ordinal values (e.g., Good = 3, Fair = 2, Poor = 1). The Spearman correlation coefficient (ρ) reveals whether raters show similar patterns in their relative assessments of subject quality, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). A higher ρ indicates that when one rater ranks a subject higher than another subject, the second rater tends to do the same. This is particularly valuable for identifying systematic differences or similarities in how pairs of raters approach the evaluation task, even if their absolute ratings differ. The Spearman correlation coefficient is calculated using the following formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

where d_i represents the difference between the ranks of each pair of observations for the two raters and n is the number of observations.

Krippendorff's Alpha (Equation 2) is employed to evaluate the inter-rater reliability across all raters simultaneously. Unlike Spearman correlation, which is limited to pairwise comparisons, Krippendorff's Alpha accounts for the agreement among multiple raters and adjusts for the probability of agreement occurring by chance. It is particularly suitable for ordinal data and provides a comprehensive measure of reliability across the entire dataset. The formula for Krippendorff's Alpha is as follows:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{2}$$

where D_o is the observed disagreement among raters and D_e is the expected disagreement by chance. An α value above 0.667 is considered acceptable for drawing meaningful conclusions from the data (Krippendorff, 2018).

By applying these two metrics, the study ensures a robust assessment of the reliability and consistency of the ratings provided by both human and AI raters, ensuring that the overall agreement is statistically significant and not merely due to chance.

4 RESULTS

The study involved six raters (five humans and one AI) who evaluated a total of 253 generated subject lines, yielding 1,518 individual ratings (253 subjects multiplied by six raters).

4.1 Data Filtering

Initial analysis revealed insufficient inter-rater reliability with Krippendorff's Alpha below 0.667, the

threshold required for drawing meaningful conclusions (Krippendorff, 2018). To achieve acceptable reliability levels, a filtering process was implemented based on inter-rater agreement across all 253 generated subject lines. Figure 2 visualizes this process through the two metrics Krippendorff's Alpha (blue line) and the proportion of retained data points (green line) at different agreement thresholds.

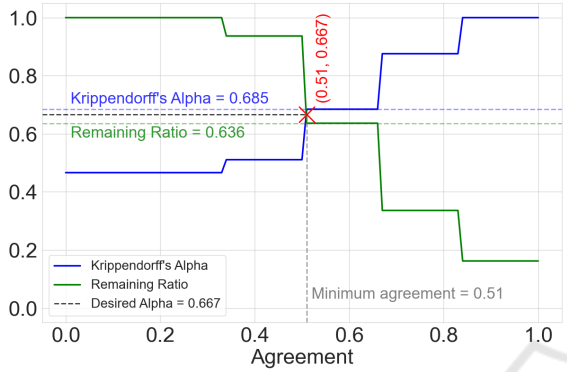

Figure 2: Illustration of the relationship between Krippendorff's Alpha (blue), remaining data ratio (green) and minimum agreement threshold (grey). The graph shows how the retained data proportion decreases as agreement thresholds rise, causing a corresponding increase in Krippendorff's Alpha values. The red cross marks where Alpha exceeds 0.667 at 51% agreement, determining the minimum agreement used for filtering.

The filtering process incrementally increases the agreement threshold (x-axis) from 0 to 1. At each threshold level, the process evaluates every generated subject line by calculating the proportion of raters who agreed on its categorisation. Subject lines that meet or exceed the current threshold are retained in the dataset, while those that fail to reach the threshold are excluded. The green line (Remaining Ratio) shows the proportion of retained subject lines at each threshold level. As the threshold increases, fewer subject lines meet the agreement criteria, causing the green line to decline. For each threshold level and its corresponding filtered dataset, Krippendorff's Alpha is recalculated (blue line). The removal of subject lines with lower agreement gradually increases the Alpha value, as shown by the rising blue line.

With six raters, the agreement thresholds create discrete steps at multiples of approximately 16.67% (100% divided by six raters), explaining the stepwise changes in both lines. The critical point occurs at a agreement threshold of 51%, where Krippendorff's Alpha crosses the desired threshold of 0.667. At this level, subject lines require agreement from more than three raters (≈ 3.06) to remain in the dataset.

After filtering, 161 out of the original 253 generated subject lines remained, resulting in a dataset reduction of 36.4%. Each model originally generated 23 subject lines. GPT-4o had the highest retention rate, with 78.3% (n=18) of its generated subject lines remaining in the dataset. SKLM Llama 3 70b Q4, Mistral 8x7b Q4 and SKLM Mixtral 8x7b Q8 each had a retention rate of 73.9% (n=17). SKLM Llama 3.1 8b Q8 retained 69.6% (n=16) of its subject lines. Meta Llama 3.1 8b Q4, GPT-3.5 Turbo and Meta Llama 3.1 8b Q8 each had a retention rate of 60.9% (n=14). SKLM Mixtral 8x7b Q4 retained 56.5% (n=13) of its subject lines after filtering. Mistral 8x7b Q8 had the lowest retention rate, with only 30.4% (n=7) of its subject lines remaining. As a consequence of the filtering process, the distribution of ratings shifted noticeably. Good ratings decreased from 532 to 370, Fair ratings from 501 to 261 and

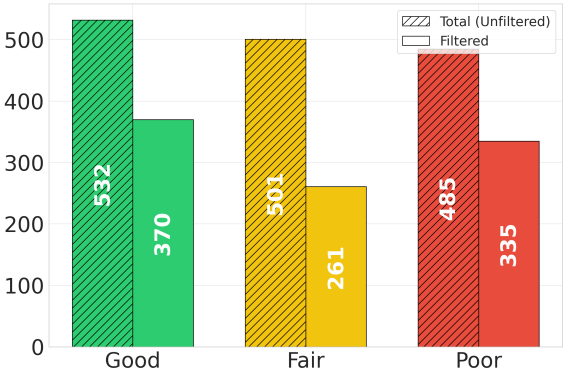

Figure 3: Distribution of ratings before (hatched bars, n = 1518) and after filtering (solid bars, n = 966). The filtering process reduced the dataset by 36.4%, with Fair ratings showing the strongest reduction (47.9%), followed by Good (30.5%) and Poor ratings (26.8%).

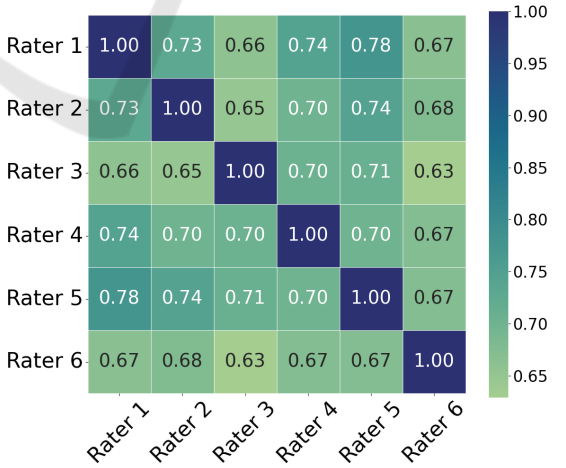

Figure 4: Pairwise Spearman correlation heatmap between the 6 Raters, after filtering. Rater 1-5 are human-raters, while Rater 6 is the AI-rater.

Poor ratings from 485 to 355, as shown in Figure 3. The reduction in the Fair ratings suggests that it was mainly in this category that the raters disagreed, this is reasonable, as the Fair category represents a middle ground where subjective interpretation is more likely to differ among raters.

This filtering process results in an increase in Krippendorff's Alpha to 0.685, surpassing the desired threshold. Figure 4 shows the Spearman correlation heatmap illustrating the pairwise agreement between the six raters after filtering. It is evident that rater 3 (human) and rater 6 (AI) have lower agreement levels with the other raters, with correlation values between 0.63 and 0.67, respectively.

4.2 Model Comparison

Following the data filtering process that ensured acceptable inter-rater reliability, the next step involved comparing the performance of the various LLMs in generating concise and relevant subject lines. Since only "Good" ratings indicate truly beneficial subject lines, while "Fair" or "Poor" ratings could potentially hinder counselling, the analysis focuses on the percentage of "Good" ratings achieved by each model. Figure 5 illustrates the aggregated ratings for all investigated models, providing a visual representation of their relative performance.

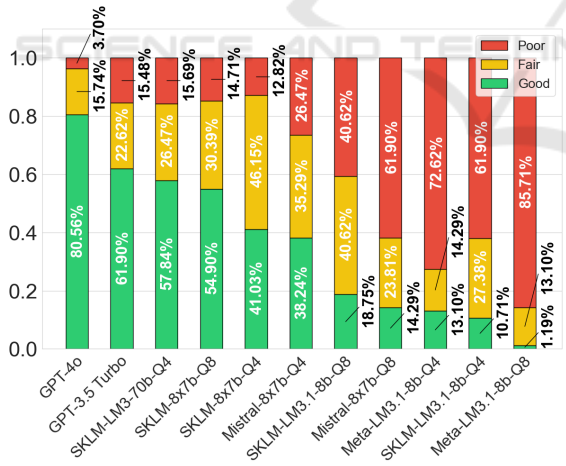

Figure 5: The distribution of filtered ratings (Good, Fair and Poor) across all evaluated models is presented. Model names are abbreviated as follows: LM3.1 (Llama 3.1), LM3 (Llama 3) and 8x7b (Mixtral 8x7b). The suffixes Q4 and Q8 denote 4-bit and 8-bit quantisation, respectively.

4.2.1 GPT vs. Mixtral vs. Llama

GPT-4o leads the way with 80.56% of its generated subject lines rated Good, making it the best performing model. GPT-3.5 Turbo also performed well, with 61.90% of its output rated Good, making it the second best model in the evaluation. While these proprietary models demonstrate superior performance, consistent with their established reputation for language generation, their use in e-mental health settings raises significant privacy concerns as sensitive data must be transmitted to external servers. The Mixtral models positioned themselves in the middle range of performance among the evaluated models. Specifically, the SKLM Mixtral 8x7b Q8 version generated 54.90% Good ratings and the SKLM Mixtral 8x7b Q4 version achieved 41.03% Good ratings. Notably, the SKLM versions consistently outperformed their standard counterparts, indicating the effectiveness of language-specific fine-tuning. In contrast, the Llama 8b models, including both the Meta and Sauerkraut variants, were generally at the lower end of the Good ratings, indicating limitations in this context. However, the SKLM Llama 3 70b Q4 model stands out as an open source model, achieving 61.90% good ratings and positioning itself as the best performing open source model.

In conclusion, the SKLM Mixtral 8x7b and Llama 3 70b Q4 models show potential and with further fine-tuning and prompt engineering, could narrow the performance gap with GPT-4o, offering valuable alternatives for online consulting applications.

4.2.2 Standard vs. German-Tuned Versions

The German-tuned Sauerkraut variants consistently outperformed their standard counterparts in both Llama and Mixtral model families, with one exception. This performance difference is particularly evident in the Mixtral models, where SKLM Mixtral 8x7b Q8 and Q4 achieved notably higher proportions of Good ratings (54.90% and 41.03%) compared to their base versions (14.29% and 38.24%). The effectiveness of language-specific fine-tuning is also demonstrated by SKLM Llama 3 70b Q4, achieving 57.84% Good ratings. Only SKLM Llama 3.1 8b Q8 performed slightly worse than its base model. Overall, these results demonstrate that fine-tuning on the target language German improves model performance in generating appropriate subject lines for German counselling communications.

4.2.3 Parameter Count and Quantisation

The analysis suggests that model size plays a crucial role in performance quality. The largest model, GPT-4o (estimated hundreds of billions of parameters), achieved the highest proportion of "Good" ratings at 80.56%. This is followed by SKLM Llama 3 70b Q4 with 57.84% "Good" ratings. The Mixtral

8x7b models, with effective 47b parameters through mixture of experts but only using 13B active parameters during inference (Jiang et al., 2024), showed varied performance. While three models achieved solid "Good" ratings between 38.24-54.90%, Mixtral 8x7b Q8 notably underperformed with only 14.29%. An interesting exception is GPT-3.5 Turbo, which despite its smaller size, with estimated only 20B parameters (Singh et al., 2023), achieved 61.90% "Good" ratings, outperforming some larger models. The smallest models, Llama 3.1 variants with 8b parameters, demonstrated the weakest performance, achieving only 1.19–18.75% "Good" ratings.

Examining the impact of quantisation, the results do not indicate a clear advantage for either higher or lower precision. For example, out of all Mistral Models the Mixtral 8x7b Q8 model produces at least Good ratings with only 14.29%, while its Q4 counterpart achieved a much better performance with 38.24% in the Good category. The Llama 3.1 8b Q8 Model even performed worst. However, no consistent trend was observed across all models to suggest that higher or lower quantisation consistently affects performance. Instead, the baseline performance of the model itself, rather than its quantisation level, appears to have the most influence on the final output quality. This result is consistent with the literature Jin et al. (2024).

These findings suggest that while model size generally correlates with improved performance, the impact of quantisation appears minimal, indicating that computational efficiency can potentially be achieved without performance degradation.

5 CONCLUSION

This study assessed 11 LLMs for generating concise and relevant subject lines in psychosocial e-mail counselling a subset of e-mental health. From 23 distinct e-mail threads, the models produced 253 subject lines. Each was evaluated by six raters — five human professionals and one AI — resulting in 1,518 ratings. After filtering for acceptable inter-rater reliability, 966 ratings were retained for analysis.

GPT-4o demonstrated superior performance with the highest proportion of Good ratings, followed by GPT-3.5 Turbo, SKLM Llama 3 70b and the Mixtral models, while smaller Llama 8b models showed limited capabilities.

The investigation yielded three key findings. First, while LLMs can generate meaningful subject lines to support counsellors in quickly grasping client issues, their current capabilities have limitations. Second, language-specific fine-tuning proves beneficial, as demonstrated by the German-tuned Sauerkraut models outperforming their base versions. Third, model size emerged as a crucial factor for performance, while quantisation showed minimal impact, suggesting that computational efficiency can be achieved without performance losses.

The study faces limitations through the relatively small number of raters and subject lines, the exclusion of full-precision and domain-specific models due to resource constraints and the closed-source nature of proprietary models restricting detailed analysis. Future research should address these limitations while aiming for a higher Krippendorff's Alpha of 0.80 to enhance reliability.

Despite these limitations, the findings suggest promising directions for practical applications. While proprietary models currently lead in effectiveness, open-source alternatives — particularly the Sauerkraut models — show potential for improvement through further fine-tuning and prompt engineering.

REFERENCES

Chiu, Y. Y. et al. (2024). A Computational Framework for Behavioral Assessment of LLM Therapists. arXiv:2401.00820.

Chung, N. C., Dyer, G., and Brocki, L. (2023). Challenges of Large Language Models for Mental Health Counseling. arXiv:2311.13857.

Dubey, A. et al. (2024). The Llama 3 Herd of Models. arXiv:2407.21783.

Fu, G. et al. (2023). Enhancing Psychological Counseling with Large Language Model: A Multifaceted Decision-Support System for Non-Professionals. arXiv:2308.15192.

Gu, Z. and Zhu, Q. (2024). MentalBlend: Enhancing Online Mental Health Support through the Integration of LLMs with Psychological Counseling Theories. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Guo, Z. et al. (2024). Large Language Models for Mental Health Applications: Systematic Review. *JMIR Mental Health*, 11:e57400.

Hackl, V. et al. (2023). Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education*, 8. Publisher: Frontiers.

Hsu, S.-L. et al. (2023). Helping the Helper: Supporting Peer Counselors via AI-Empowered Practice and Feedback. arXiv:2305.08982.

Jiang, A. Q. et al. (2024). Mixtral of Experts. Version Number: 1.

Jin, R. et al. (2024). A Comprehensive Evaluation of Quantization Strategies for Large Language Models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics:*

ACL 2024, pages 12186–12215, Bangkok, Thailand. Association for Computational Linguistics.

Khan, M. P. and O'Sullivan, E. D. (2024). A comparison of the diagnostic ability of large language models in challenging clinical cases. *Frontiers in Artificial Intelligence*, 7:1379297.

Kocbek, P. et al. (2022). Generating Extremely Short Summaries from the Scientific Literature to Support Decisions in Primary Healthcare: A Human Evaluation Study. In Michalowski, M., Abidi, S. S. R., and Abidi, S., editors, *Artificial Intelligence in Medicine*, pages 373–382, Cham. Springer International Publishing.

Koutsouleris, N. et al. (2022). From promise to practice: towards the realisation of AI-informed mental health care. *The Lancet Digital Health*, 4(11):e829–e840. Publisher: Elsevier.

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications. Google-Books-ID: nE1aDwAAQBAJ.

Lai, T. et al. (2024). Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs). *BioMedInformatics*, 4(1):8–33. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Lawrence, H. R. et al. (2024). The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Mental Health*, 11(1):e59479. Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.

Li, A. et al. (2024). Understanding the Therapeutic Relationship between Counselors and Clients in Online Text-based Counseling using LLMs. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1280–1303, Miami, Florida, USA. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, J. M. et al. (2023). ChatCounselor: A Large Language Models for Mental Health Support. arXiv:2309.15461.

Mirzakhmedova, N. et al. (2024). Are Large Language Models Reliable Argument Quality Annotators? In Cimiano, P. et al., editors, *Robust Argumentation Machines*, pages 129–146, Cham. Springer Nature Switzerland.

Papineni, K. et al. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rao, A., Aithal, S., and Singh, S. (2024). Single-Document Abstractive Text Summarization: A Systematic Literature Review. *ACM Comput. Surv.*, 57(3):60:1–60:37.

Rudolph, E., Engert, N., and Albrecht, J. (2024). An AI-Based Virtual Client for Educational Role-Playing in the Training of Online Counselors:. In *Proceedings of the 16th International Conference on Computer Supported Education*, pages 108–117, Angers, France. SCITEPRESS - Science and Technology Publications.

Shahriar, S. et al. (2024). Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. *Applied Sciences*, 14(17):7782. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

Singh, M. et al. (2023). CodeFusion: A Pre-trained Diffusion Model for Code Generation. arXiv:2310.17680 version: 3.

Tam, T. Y. C. et al. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7(1):1–20. Publisher: Nature Publishing Group.

Temsah, M.-H. et al. (2024). OpenAI o1-Preview vs. ChatGPT in Healthcare: A New Frontier in Medical AI Reasoning. *Cureus*, 16(10):e70640.

Vowels, L. M., Francois-Walcott, R. R. R., and Darwiche, J. (2024). AI in relationship counselling: Evaluating ChatGPT's therapeutic capabilities in providing relationship advice. *Computers in Human Behavior: Artificial Humans*, 2(2):100078.

Wu, Y. et al. (2024). Less is More for Long Document Summary Evaluation by LLMs. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 330–343, St. Julian's, Malta. Association for Computational Linguistics.

Xu, X. et al. (2024). Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1):31:1–31:32.

Yin, Y.-J., Chen, B.-Y., and Chen, B. (2024). A Novel LLM-based Two-stage Summarization Approach for Long Dialogues. arXiv:2410.06520.

Zhang, H., Yu, P. S., and Zhang, J. (2024a). A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models. arXiv:2406.11289.

Zhang, R. and Tetreault, J. (2019). This Email Could Save Your Life: Introducing the Task of Email Subject Line Generation. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 446–456, Florence, Italy. Association for Computational Linguistics.

Zhang, T. et al. (2019). BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhang, T. et al. (2024b). Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.