Team Conversational AI: Introducing Effervesce

Erjon Skënderi
University of Helsinki
erjon.skenderi@helsinki.fi

Salla-Maaria Laaksonen
University of Helsinki
salla.laaksonen@helsinki.fi

Jukka Huhtamäki
Tampere University
jukka.huhtamaki@tuni.fi

Abstract

Group conversational AI, especially within digital workspaces, could potentially play a crucial role in enhancing organizational communication. This paper introduces Effervesce, a Large Language Model (LLM) powered group conversational bot integrated into a multi-user Slack environment. Unlike conventional conversational AI applications that are designed for one-to-one interactions, our bot addresses the challenges of facilitating multi-actor conversations. We first evaluated multiple open-source LLMs on a dataset of 1.6k group conversation messages. We then fine-tuned the best performing model using a Parameter Efficient Fine-Tuning technique to better align Effervesce with multi-actor conversation settings. Evaluation through workshops with 40 participants indicates positive impacts on communication dynamics, although areas for further improvement were identified. Our findings highlight the potential of Effervesce in enhancing group communication, with future work aimed at refining the bot's capabilities based on user feedback.

1 Introduction

In the era of digital workspaces, organizations are increasingly communicating using different online tools that facilitate interaction and collaboration on the level of the entire organization or in teams (Sinclaire and Vogus, 2011, e.g.,). Recent studies have highlighted the importance of online collaboration software (OCSs), particularly for teamwork, and even more specifically, for distributed, virtual teams (Gilson et al., 2015; Ford et al., 2017; Laitinen and Valo, 2018). Organizational virtual teams are relatively small, task-oriented groups of individuals who are often physically distributed to multiple locations nation- or worldwide, and mostly work technology-mediated toward a common goal (Berry, 2011; Lipnack and Stamps, 2008). When shared physical premises are lacking, the importance of online collaboration software becomes even more evident: It becomes the site where both work-related and relational team processes take place (e.g., Laitinen and Valo, 2018; Gibbs et al., 2008; Laitinen et al., 2021). OCSs, thus, facilitate various team processes across temporal and physical boundaries, as well as allow team members to get to know each other by providing a shared platform for the team to socialize on (Stoeckli et al., 2020).

Researchers have extensively discussed how technology integrates into organizational life: it shapes social action of organization members and technology itself is also shaped through people using it (Leonardi and Barley, 2010; Orlikowski, 2007) . During the past few years, the development of Large Language Models (LLMs) and chatbots built using them has radically changed the type of technologies used in organizational communication. Through the advances of communicative AI, the role of technology develops from a tool that *affords* communication to a tool that *participates* in human interaction. In communication scholarship, the term "communicative AI" has been coined to refer to devices, applications, and algorithms capable of communicating in natural language and adapting to real-life conversational situations (Guzman and Lewis, 2020; Jones, 2014). In computer science, these applications have been discussed under the term conversational AI (e.g., Kulkarni et al., 2019; McTear, 2022). The future projections of companies such as OpenAI even suggest that nonhuman conversational agents could soon be indistinguishable from humans (B., 2023).

In this study, we start from the premise that communicative AI applications, communication tools that are enhanced with LLMs and Generative AI (GenAI), could play a critical role in facilitating effective group conversations. Traditional conversational AI applications are predominantly designed for one-to-one interactions in the form of chat [1,2], which also applies to the most widely used conver-

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 502–513
May 3-4, 2025 ©2025 Association for Computational Linguistics

sational AI tools such as ChatGPT or Microsoft Copilot also used in a professional context. To facilitate team conversation and collaboration, the conversational AI should be able to take part in group conversations. This generates a need for models and applications that can support many-to-many conversations. Such an AI application could enter the team OCS with its own account and join the conversation almost as a team member. In addition, it should be able to read the flow of conversation and adapt to the language style of the team.

In this work, we introduce Effervesce, an LLM-powered group conversational bot operating on Slack designed to integrate into group conversations and engage as an AI team member in the organization's digital workspace. To power our chatbot, or more accurately a socialbot (Gehl and Bakardjieva, 2016), we evaluate various open source models that provide us with robust version control and help address data privacy concerns. Increasingly, alternative open source LLMs are being introduced in multiple recent works, including Llama (Touvron et al., 2023b; Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), and Qwen (Bai et al., 2023). We created a group-conversational dataset from the 1,608 messages posted on a Slack channel of a single team. In our preliminary evaluations of various open-sourced LLMs with our group conversational dataset, we observed that such models struggled to capture the language style and structure of the conversational context. We selected the best-performing model, a fine-tuned variant of Mistral-7B, to power Effervesce.

We addressed the identified issues with context understanding by fine-tuning the selected LLM. We acknowledge that there are substantial costs and environmental implications associated with training and fine-tuning such large machine learning models (Jiang et al., 2024). To minimize these effects, we experimented with a Parameter-Efficient Fine-Tuning (PEFT) technique, known as QLoRA (Dettmers et al., 2023). This method allowed us to update only a small fraction of the total 7+ billion parameters while maintaining the pre-trained model's performance. Our fine-tuned version of the model managed to learn from the training data while maintaining a good generalization level.

To assess the fine-tuned Effervesce, we conducted a qualitative evaluation through 10 workshop sessions, involving 40 participants in total. The feedback was useful to guide future improvement in our approach and, among others, indicated that while the bot demonstrated improved conversation and context awareness, it responded too quickly and provided long detailed responses.

The contributions of this work are as follows.

i. We present *Effervesce* as a group conversation chatbot integrated with Slack and designed to engage with real-time multi-actor conversations.

ii. We document the dataset construction from a team's digital conversation messages, posted on Slack.

iii. We evaluate the performance of multiple pre-trained open-source LLMs on our multi-user conversation dataset.

iv. We employ and document an efficient QLoRA-based fine-tuning approach for an LLM powering our group conversational chatbot.

v. We conduct a human-centric evaluation of Effervesce through workshops with diverse groups of users. The feedback provides insights for future improvements of our chatbot.

In the following section, we summarize existing research in group conversational AI systems, technicalities and costs concerning the pre-training and fine-tuning of LLMs. In Section 3, we discuss the methodology of this work, presenting details on our dataset, LLM evaluation, and the fine-tuning approach that we employ. We describe the experiment and disseminate the results in Section 4, while in Section 5 we provide a discussion of the results and conclude this work. Lastly, in Section 6 we list the future work leads that emerge from this research.

2 Background and Related Work

Communication in organizations has increasingly shifted to online collaboration software, where teams collaborate in shared systems. Nowadays, human users on such systems are increasingly accompanied by different AI tools designed to help their workflows, knowledge management, and communication. In general, the introduction of GenAI tools in work life is expected to shape agency and action in knowledge work: routines, processes, and also professional interactions (Ramaul et al., 2024; Retkowsky et al., 2024) Previous studies on conversational bots in organizations show how AI agents can mediate human interaction and facilitate knowledge sharing (Chiang et al., 2024; Boyd et al., 2020; Ramjee et al., 2024). Only a few studies have focused on bots that take part in group conversations (Laitinen et al., 2021; Meske and Amojo, 2018), but these bots have represented pre-GenAI era bots with quite simple communication capabilities. However, most conversational AI systems used and studied so far have been applications that enable one-to-one or user-assistant interactions (Liu et al., 2023; Touvron et al., 2023a; Jiang et al., 2023; Serban et al., 2015). Consequently, research has focused on communication processes such as simple question-answering or knowledge sharing, without exploring the application of LLMs in real-time group conversations.

More recent works analyze the value of contextual understanding in group conversation settings, particularly relevant in online digital platforms like Community Question and Answering, Slack, and Reddit (Boyd et al., 2020), where multiple members can engage in conversations across different channels, threads, and topics. Various technical and design challenges arise when employing such multi-user conversational AI systems. Most notably, the conversation AI system should be able to follow the structure of the conversation and take into account that there are multiple participants involved. These challenges require AI models that keep track of dynamic conversations, recognize multiple speakers, and follow the discussion's context. Transformer-based architectures (Vaswani et al., 2017; Devlin et al., 2019) proved that contextual embeddings can capture special language features from text giving shape to the Natural Language Processing (NLP) research landscape. This attribute has enabled the development of Large Language Models (LLM), which have shown superior performance on a wide range of benchmark tasks. Early works like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) proved the power of pre-training deep transformer-based models on massive textual datasets to extract improved features from written language. BERT was followed by other models like OpenAI's Generative Pre-trained Transformer (GPT) models (Radford et al., 2019; Brown et al., 2020), which demonstrated superior few-shot learning capabilities. More recent models such as Google's Language Model for Dialogue Applications (LaMDA) (Thoppilan et al., 2022), Gemini (Team, 2024), and Deepseek-R1 (DeepSeek-AI, 2025) have become foundational models for various products and applications, including intelligent chatbots.

These models often contain billions or even trillions of parameters, posing significant challenges for implementation. Training and deploying such large models requires vast amounts of computational resources and energy, making them expensive and less accessible. Fine-tuning these models for specific tasks can also be computationally intensive (Jiang et al., 2024). Recent works have introduced various parameter-efficient fine-tuning (PEFT) techniques as solutions to address these challenges. Methods like Low-Rank Adaption (LoRA) (Hu et al., 2021) and Quantized LoRA (QLoRA) (Dettmers et al., 2023) provide alternative efficient techniques to fine-tune pre-trained LLMs for specific data and application contexts, by enabling training only on a small fraction of the model's parameters.

The challenges of enabling a chatbot to adopt different roles in multi-user conversations have been identified and explored also by Boyd et al., 2020, who introduced an augmented and fine-tuned GPT-2 model (Radford et al., 2019), which emulates the persona of a target actor based on previous conversations they engaged with. Their large-scale Reddit dataset of 10.3 million conversations enabled fine-tuning without employing parameter-efficient techniques. However, such an approach can be expensive or not feasible, especially for smaller organizations or limited datasets.

3 Methodology

In this section, we describe our approach to building and evaluating Effervesce, our Slack-based group conversation bot. First, we describe how we constructed the group conversation dataset from real Slack messages. Next, we explain the process of evaluating, selecting, and efficiently fine-tuning open-source LLMs, to power our chatbot. Finally, we describe how we evaluated the bot's performance based on the quantitative metrics and the qualitative feedback we received from human users who interacted with our bot in workshop settings.

3.1 Dataset Construction

We compiled a dataset of 1,608 Slack messages, consisting of real-world day-to-day interactions between 7 members of a research group. The data set was filtered to include only English messages. We used *"###"* as a standard annotation to define roles or users within our training data. As such, each message was annotated following a specific template that consists of 3 parts: *"###"* + *USERNAME:* + *MESSAGE*. To accommodate the model learning the language style and structure from the training data, we formatted the data as follows.

```
{
  "context": "###YOU: Raw data, om nom nom!\n"
             "###Jukka: There is no raw data, I
              ↪  mind you!\n",
  "target": "###YOU: Raw data is an oxymoron.
  ↪ - L. Gitelman ### END"
}
```

Listing 1: Training Data Sample

Each data point consists of the *context* part, which the model uses as a seed to start generating text, and the *target*, which corresponds to the desired output, which the model will attempt to learn. This approach allows the model to learn the many-to-many structure of real-time group conversations by capturing conversation flow across multiple roles. During our fine-tuning experiment, we kept 321 data points as testing data, and the rest was used to fine-tune a selected LLM.

3.2 Model Selection and Fine-Tuning

To address the challenges posed by multi-actor conversation data, we experimented with top-performing open source LLMs that were available at the time when our experiment was conducted. Specifically, we evaluate the performance of four Llama-2 models (Touvron et al., 2023b), and two variations of the Mistral 7B model (Jiang et al., 2023), namely Mistral-7B-v0.1 and Mistral-7B-Instruct-v0.1. The models we tested during the evaluation and selection phase were all in half-precision floating point (FP16) format, non-quantized versions.

To fine-tune the best-performing foundation model, we employ a Parameter-Efficient Fine-Tuning (PEFT) technique known as QLoRA (Dettmers et al., 2023). The authors of this approach claim it facilitates fine-tuning of a quantized 4-bit model without sacrificing the performance. First, a high-precision technique is employed to quantize a pre-trained model to 4-bit, then a set of Low-Rank Adapter (LoRA) weights are introduced, based on the strategy introduced by Hu et al., 2021.

3.3 Evaluation

We evaluate Effervesce using two methods. First, we quantitatively measure the performance of the selected language models using BLEU scores and perplexity. Second, we perform a qualitative analysis based on feedback from user workshops to assess the bot's interaction and overall performance.

3.3.1 Metrics for Language Models

We evaluate the performance of the LLMs we employ using two metrics: Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and perplexity.

BLEU is an n-gram-based metric for the syntactic similarity between the generated text and target text, provided as ground truth. This technique is typically applied to Machine Translation problems, however, its popularity has increased among various applications on natural language generation systems (Sai et al., 2023). The range of BLEU scores can be interpreted as a percentage, where a score of 100% indicates a perfect syntactical match between the two texts being compared.

The second metric that we use, perplexity, is a standard metric that measures how well a language model predicts a sequence of words or tokens from a given text (Meister and Cotterell, 2021). A lower perplexity score indicates that the model is less "perplexed", and more accurate at predicting the next tokens of a text. High perplexity score suggests that the generative model is struggling to predict the next tokens comprising a certain target text.

3.3.2 Human Feedback Analysis

To implement Effervesce as a group conversational bot, we integrated with Slack to listen for new messages on a specified channel and generate real-time responses based on the discussion.

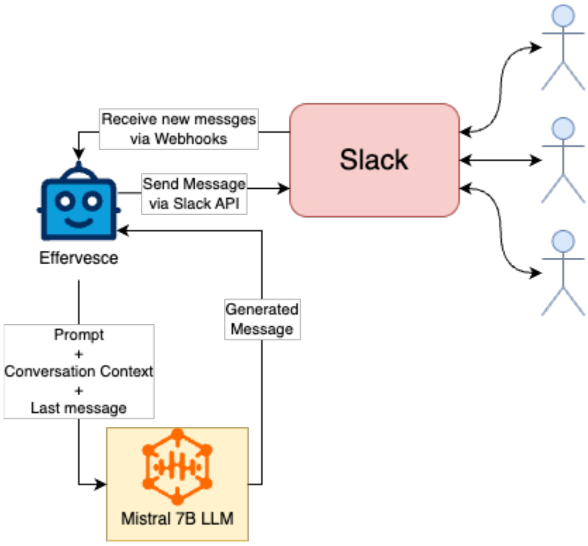

Figure 1: Effervesce workflow diagram.

The workflow diagram in Figure 1 shows how the system is set up to handle interactions between Effervesce and users through the Slack API, and the LLM as the engine for response generation. First, new messages are received from Slack using the webhooks functionality. Next, a textual prompt is combined with the conversation history and then forwarded to the LLM which generates the response. The response is then posted as a reply on behalf of the bot, through the Slack API. An example of the prompt we use is as follows.

```
"You are Effervesce, a collaborative team bot
designed to enhance discussions and
brainstorming. Your goal is to keep
conversations focused, productive, and
creative. Provide concise, relevant responses
to encourage collaboration and new ideas while
ensuring the team stays on topic and on time.

Conversation context will follow this format:

user1: 'message'
Effervesce: 'reply'
user2: 'message'
Effervesce: 'reply'

Stick to the context, foster teamwork, and
maintain brevity in your replies."
```

Listing 2: Effervesce's prompt.

The qualitative evaluation of Effervesce was carried out as part of a workshop setting in which human participants were invited to test the prototype bot. We ran 10 workshops with 40 participants in total. The participants represented communication professionals, IT consultants, forest industry, as well as university students in communication/language studies and IT management. In the workshops, the participants were first asked to engage in a team discussion with a creative task so that the bot was taking part in the conversation. Afterward, the groups were asked to jointly reflect on the experience and assess how the bot worked, how it impacted the conversation, and how would they wish to change the bot. These group discussions were recorded and transcribed, and the recordings were qualitatively analyzed to map how participants assessed the bot's performance in a group conversation setting.

4 Experiment and Results

We present our experiments and findings on evaluating a set of LLMs on group conversational data and fine-tuning and evaluating an LLM to power our group chatbot. First, we describe the experimental setup. Then, we organize our results in two groups: 1) evaluating different LLMs with our group conversational data, and 2) Fine-tuning and Qualitative Assessment of Effervesce.

4.1 Experimental Setup

For this experiment, we employ a machine equipped with two NVIDIA Tesla V100 PCIe 16GB GPUs. We run our evaluations, fine-tuning, and deployment using Python, and use HuggingFace's *transformers* library to load and interact with the selected LLMs. We use cross-entropy loss function during the fine-tuning process with QLoRA. Out of 7.28 billion total parameters, only 42 million (0.58%) were set to be trainable. We set some of the key parameters to the following values: 1) LoRA: *rank=16, alpha=64, dropout=0.1;* 2) Fine-tuning: *learning_rate=2e-4, batch_size=4, gradient_acc=4;*

During our qualitative assessment through workshops, we deployed Effervesce as a Flask-based web application. A web interface was made accessible to us authors, providing system information and implementing a probability slider functionality to adjust how frequently the bot engaged in conversations. By default, this parameter was set to 60%, and the chatbot would reply automatically to 60% of new messages unless it was specifically mentioned in a conversation as *@EffervesceBot*.

4.2 Experiment Part 1: LLM Evaluation in Group Conversation Context

We measure the performance of the six large language models that were selected to be evaluated in our group conversation dataset. The average perplexity and BLEU scores achieved from all models are provided in Table 1.

In our experiment, both the pre-trained Llama-2 models and the pre-trained *Mistral-7B-v0.1* models achieve lower perplexity scores compared to their corresponding fine-tuned versions. These versions have been explicitly fine-tuned to follow instructions or answer questions in a one-to-one fashion. While the perplexity scores are high overall, the difference among these two groups of models is significant.

Mistral-7B-Instruct-v0.1 model achieved the highest BLEU score of 9.27%, indicating that the responses generated by this model were the most syntactically similar to the reference text. While its perplexity score of 42.30 was worse than the scores achieved from the pre-trained models, this difference is argued in previous research (Meister and Cotterell, 2021; Sai et al., 2023) which shows that fine-tuned natural language generation models often optimize for the language style and content alignment over statistical prediction. In their work, Jiang et al., 2023, highlighted that "Mistral-7B-v0.1 outperforms Llama 2 13B on multiple natural language generation benchmarks". In our experiments, we were able to validate this indicated performance improvement in our data context as well.

Beyond the quantitative evaluation, we also interacted with the bot directly, while powered by this specific model. Subjectively, the responses generated by *Mistral-7B-Instruct-v0.1* followed a more natural conversation flow, were more aware of the conversation context, and followed the directions given through the prompt better.

4.3 Experiment Part 2: Fine-tuning and Qualitative Assessment of Effervesce

The Mistral-7B-Instruct-v0.1 was selected as the LLM to power our group conversational chatbot. We fine-tuned the model using our group conversation dataset to align it with the language style, vocabulary, and multi-actor configuration.

Figure 2 shows how the model's perplexity decreased during the fine-tuning epochs. Initially, the model started with perplexity varying between 32-55, and then gradually dropped closer to 3 during the training. The error rate decreased quickly due to the amount of training data available, and the relatively small amount of trainable weights introduced by the QLoRA technique.

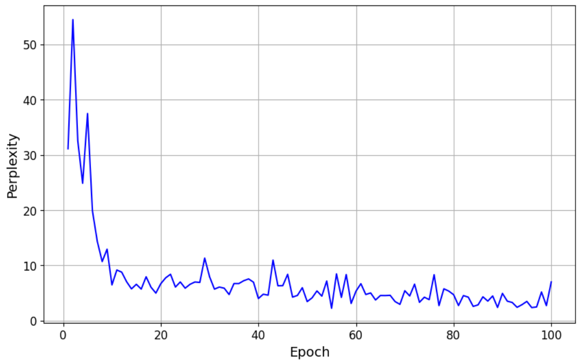

Figure 2: Perplexity over 100 epochs.

We measured a 7.8 average perplexity of the fine-tuned model when evaluated on the 321 test data points. This score is larger than the best score achieved on the training set. However, this indicates the model did not overfit with the training set, regardless of the small amount of training data.

To evaluate the performance of Effervesce in real-world conversations, we conducted 10 workshop sessions where participants interacted with the bot in group configurations. The qualitative feedback that we received can be grouped as follows. **1) The bot was too active and quick to respond.** During the first several workshop sessions, Effervesce was set to reply to every message. This caused it to dominate the conversations, resulting in causing some of the other participants to not engage. For the following workshop sessions, we introduced a probability-of-response slider which was controlled through a web interface.

2) Responses were too long and too detailed. The bot provided too many suggestions, often in the form of bullet points, making its replies difficult to follow.

3) The language style of the bot seemed overly friendly and informal. The bot used too many emojis and was overly positive, which some participants did not find natural in a professional environment.

4) The bot made mistakes. Effervesce occasionally used the wrong names while referring to the participants. It would either make grammatical typos or refer to a different person in the conversation.

5) The bot failed to offer critical feedback.

Type	Model	Perplexity	BLEU(%)
Pre-trained	Llama-2 7B	29.48	2.96
	Llama-2 13B	29.40	3.04
	Mistral-7B-v0.1	**29.18**	3.49
Fine-tuned	Llama-2 7B-chat	62.09	5.19
	Llama-2 13B-chat	55.40	6.84
	Mistral-7B-Instruct-v0.1	42.30	**9.27**

Table 1: Perplexity (lower the better) and BLEU(%) (higher the better) on our Slack group conversation dataset.

By design, the bot was prompted to be supportive and encouraging. Some participants did not find it useful when having brainstorming sessions.

These findings indicate that the fine-tuned Effervesce was perceived as dynamic, but also that its participation could disrupt the natural group interaction.

5 Discussion and Conclusion

In this work, we explored how Effervesce, our group conversational chatbot, integrated with Slack and designed to engage in real-time multi-actor conversations. We evaluated multiple open-source LLMs, fine-tuned *Mistral-7B-Instruct-v0.1* model using the QLoRA technique, and evaluated Effervesce's performance through quantitative metrics and qualitative user feedback.

Pre-trained models achieved lower perplexity scores, compared to their fine-tuned counterparts, when evaluated in our group conversational dataset. However, these foundation models performed worse based on BLEU scores, suggesting their lack of alignment with the language style in the group conversation. Fine-tuned models improved BLEU scores consistently, but performed worse on the perplexity metric. Given these findings, we selected *Mistral-7B-Instruct-v0.1* model for further fine-tuning and powering Effervesce due to its better performance in instruction-following, group context understanding, and higher response quality perceived by humans.

Perplexity decreased during the fine-tuning process, indicating that the model managed to learn the language style and patterns from our specific application data context. We evaluated the fine-tuned model on our test set, and it achieved an average perplexity score of 7.8, indicating the model did not overfit.

Through our qualitative evaluation, we received feedback regarding Effervesce's performance in group conversation settings. The bot was perceived as too active during the first interactions, disrupting the flow of the conversation. We introduced a response probability parameter in the system, which helped to improve this concern for the following workshops.

Some users found Effervesce's responses too long and overwhelming. We received feedback indicating the bot's language tone was found to be too friendly, using a lot of emojis, and informal language considering the professional context of evaluation. Our chatbot also made mistakes when referring to users participating in the conversation. Mistakes were in the form of typos and complete misses. Some users felt the bot did not provide critical feedback when asked to facilitate their brainstorming session.

Effervesce demonstrates the potential of LLM-powered multi-actor chatbots in digital workspaces to enhance group communication dynamics in organizations. Fine-tuning improved its performance and alignment with the group conversation structure and dynamics. Nevertheless, the user feedback pointed out further challenges that the bot faces. Addressing the identified issues is crucial for further investigating how to make group conversational AI more effective.

Our work contributes to the growing research field of LLM-powered multi-actor group conversation chatbots through the insights we provided regarding the LLM fine-tuning, and practical integration and deployment process.

6 Future Work

Effervesce demonstrated its potential to facilitate group conversations. However, several areas require further investigation. Future work will focus on improving the dataset quality and size, exploring recent open-source LLM alternatives, and enhancing Effervesce's behavior based on the evaluation outcomes of this work. Our goal is to further research the bot's turn-taking functionality, enhancing the response strategy by integrating various recently introduced functionalities, like tool-calling (Shen, 2024) and Retrieval Augmented Generation (RAG) (Lewis et al., 2020).

Larger and more diverse training datasets could potentially help Effervesce better generalize and align with the structure of group conversations. Such an improvement would have a positive impact on reducing hallucinations when referring to other users by name. Additionally, future work could explore how fine-tuning and evaluating the bot with data originating directly from the team it is interacting with impacts the bot's performance.

Numerous effective open-source LLMs have been published recently. Our investigation can be extended by comparing the performance in group conversation settings of alternative models such as Qwen (Bai et al., 2023), DeepSeek (DeepSeek-AI, 2025), and Llama 3.1 (Grattafiori et al., 2024).

Future works can explore different fine-tuning strategies, including fine-tuning with alternative quantization techniques, and investigate how implementing other PEFT techniques could impact the LLM's performance.

Various strategies can be employed to improve Effervesce's behavior in conversations. Effervesce currently responds to new messages based on a hard-coded probability parameter. Future work can focus on implementing alternative turn-taking prediction mechanisms, so the bot knows when to engage in a conversation and when to remain silent. This could optimize the response length and language style to make interactions feel more natural and professional on the other users' side. In future versions of our bot, we will consider implementing features and checks to ensure the bot does not overwhelm human team members and facilitates a balanced participation of all.

Lastly, future work can also test several features to improve Effervesce's utility in work or professional environments. We will implement function or tool-calling capabilities, which will enable the bot to interact with external tools and databases in real-time. In addition, advanced context retrieval techniques like RAG could be implemented to improve the bot's interaction quality.

Limitations

Our study has several limitations, listed as follows.

Training Dataset Size and Context. The fine-tuning dataset consists of 1,608 Slack messages from a single research group. LLMs trained with this data result in limited generalization capabilities for other teams and contexts.

Fine-tuning Technique. While using an efficient technique like QLoRA to fine-tune our bot costs less, it also restricts how much the model could learn with full fine-tuning.

Evaluation Metrics. Perplexity and BLEU scores do not consider the conversation flow and engagement level in multi-actor conversations.

Turn-Taking. Effervesce doesn't regulate its engagement in a conversation, disrupting the natural conversation flow, and affecting the user's perception of the bot.

References

Tamim B. 2023. Chatgpt-powered ai legal assistant launches and brings along fear. *Interesting Engineering*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, Zhu Qwen Team, and Alibaba Group. 2023. Qwen Technical Report.

Gregory R Berry. 2011. A cross-disciplinary literature review: Examining trust on virtual teams. *Performance Improvement Quarterly*, 24(3):9–28.

Alex Boyd, Raul Puri, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Large scale multi-actor generative dialog modeling. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 66–84.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 2020-December.

Chun Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision

Making through LLM-Powered Devil's Advocate. *ACM International Conference Proceeding Series*, 17:103–119.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186.

Robert C Ford, Ronald F Piccolo, and Loren R Ford. 2017. Strategies for building effective virtual teams: Trust is key. *Business horizons*, 60(1):25–34.

Robert W. Gehl and Maria Bakardjieva. 2016. *Socialbots and their friends : digital media and the automation of sociality*. Routledge.

Jennifer L. Gibbs, Dina Nekrassova, Svetlana V. Grushina, and Sally Abdul Wahab. 2008. Reconceptualizing virtual teaming from a constitutive perspective review, redirection, and research agenda. *Annals of the International Communication Association*, 32:187–229.

Lucy L. Gilson, M. Travis Maynard, Nicole C. Jones Young, Matti Vartiainen, and Marko Hakonen. 2015. Virtual teams research. *Journal of Management*, 41:1313–1337.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models.

Andrea L. Guzman and Seth C. Lewis. 2020. Artificial intelligence and communication: A human–machine communication research agenda. *New Media and Society*, 22:70–86.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR 2022 - 10th International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Peng Jiang, Christian Sonne, Wangliang Li, Fengqi You, and Siming You. 2024. Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots. *Engineering*, 40:202–210.

Steve Jones. 2014. People , things , memory and human-machine communication. *International Journal of Media & Cultural Politics*, 10:245–258.

Pradnya Kulkarni, Ameya Mahabaleshwarkar, Mrunalini Kulkarni, Nachiket Sirsikar, and Kunal Gadgil. 2019. Conversational ai: An overview of methodologies, applications & future scope. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–7.

Kaisa Laitinen, Salla-Maaria Laaksonen, and Minna Koivula. 2021. Slacking with the bot: Programmable social bot in virtual team interaction. *Journal of Computer-Mediated Communication*, 26:343–361.

Kaisa Laitinen and Maarit Valo. 2018. Meanings of communication technology in virtual team meetings: Framing technology-related interaction. *International Journal of Human-Computer Studies*, 111:12–22.

P. M. Leonardi and S. R. Barley. 2010. What's under construction here? social action, materiality, and power in constructivist studies of technology and organizing. *The Academy of Management Annals*, 4:1–51.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen Tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 2020-December.

Jessica Lipnack and Jeffrey Stamps. 2008. *Virtual teams: People working across boundaries with technology*. John Wiley & Sons.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models. *Meta-Radiology*, 1(2).

Michael McTear. 2022. *Conversational ai: Dialogue systems, conversational agents, and chatbots*. Springer Nature.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 5328–5339.

Christian Meske and Ireti Amojo. 2018. Social bots as initiators of human interaction in enterprise social networks. *ACIS 2018 - 29th Australasian Conference on Information Systems*, pages 1–7.

W. J. Orlikowski. 2007. Sociomaterial practices: Exploring technology at work. *Organization Studies*, 28:1435–1448.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.

Laavanya Ramaul, Paavo Ritala, and Mika Ruokonen. 2024. Creational and conversational AI affordances: How the new breed of chatbots is revolutionizing knowledge industries. *Business Horizons*, 67(5):615–627.

Pragnya Ramjee, Bhuvan Sachdeva, Satvik Golechha, Shreyas Kulkarni, Geeta Fulari, Kaushik Murali, and Mohit Jain. 2024. CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients.

Jana Retkowsky, Ella Hafermalz, and Marleen Huysman. 2024. Managing a ChatGPT-empowered workforce: Understanding its affordances and side effects. *Business Horizons*, 67(5):511–523.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Computing Surveys*, 55(2):39.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pages 3776–3783.

Zhuocheng Shen. 2024. LLM With Tools: A Survey.

Jollean K. Sinclaire and Clinton E. Vogus. 2011. Adoption of social networking sites: An exploratory adaptive structuration perspective for global organizations. *Information Technology and Management*, 12(4):293–314.

Emanuel Stoeckli, Christian Dremel, Falk Uebernickel, and Walter Brenner. 2020. How affordances of chatbots cross the chasm between social and traditional enterprise systems. *Electronic Markets*, 30:369–403.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee Huaixiu, Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel, Morris Tulsee, Doshi Renelito, Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le Google. 2022. LaMDA: Language Models for Dialog Applications.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael, Smith Ranjan, Subramanian Xiaoqing, Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 5999–6009. Neural information processing systems foundation.