TaskDiff: A Similarity Metric for Task-Oriented Conversations

Ankita Bhaumik[†], **Praveen Venkateswaran**[*], **Yara Rizk**[*], **Vatche Isahagian**[*]
[†] Rensselaer Polytechnic Institute, Troy, New York
[*] IBM Research
bhauma@rpi.edu
{praveen.venkateswaran, yara.rizk, vatchei}@ibm.com

Abstract

The popularity of conversational digital assistants has resulted in the availability of large amounts of conversational data which can be utilized for improved user experience and personalized response generation. Building these assistants using popular large language models like ChatGPT also require additional emphasis on prompt engineering and evaluation methods. Textual similarity metrics are a key ingredient for such analysis and evaluations. While many similarity metrics have been proposed in the literature, they have not proven effective for task-oriented conversations as they do not take advantage of unique conversational features. To address this gap, we present *TaskDiff*, a novel conversational similarity metric that utilizes different dialogue components (utterances, intents, and slots) and their distributions to compute similarity. Extensive experimental evaluation of *TaskDiff* on a benchmark dataset demonstrates its superior performance and improved robustness over other related approaches.

1 Introduction

Task-oriented conversational assistants have become increasingly popular in multiple industries enabling users to perform tasks such as travel reservations, banking transactions, online shopping, etc., through multi-turn conversations. The increased use of these assistants has led to the availability of valuable user-assistant conversation logs (Budzianowski et al., 2018; Andreas et al., 2020), prompting efforts to extract insights from them.

A key aspect of such conversational analytics is identifying similarities and dissimilarities between conversations. This will enable developers to improve the user-experience including personalized response generation, next-action recommendations, and information retrieval (Yaeli et al., 2022; Bag et al., 2019; Gao et al., 2020; Li et al., 2022). The popularity of large language models like ChatGPT and Llama 2 (Touvron et al., 2023) has resulted in a race to create custom task-oriented conversational assistants in enterprise domains like finance and retail (Wu et al., 2023). However, evaluating these assistants has become an important challenge and requires effective metrics that can measure their performance across similar user-assistant conversations.

Measuring semantic textual similarity has been extensively studied for textual sources like documents, social media, transcripts, etc. However, there has been limited prior work studying similarity in task-oriented conversation settings (Appel et al., 2018; Lavi et al., 2021). Most approaches leverage popular word embeddings like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) or pre-trained models like Universal Sentence Encoder (Cer et al., 2018), Sentence-BERT (Reimers and Gurevych, 2019) to obtain vector representations of utterances, and then use distance-based approaches such as cosine and edit-distance to compute the similarity between text snippets.

While such approaches can identify semantic relationships between texts, task-oriented conversations present several challenges that limit their effectiveness. Firstly, they consist of distinct components – intents, slots, and utterances – that impact the similarity and overlap between conversations. For instance, users can have different objectives (e.g., booking travel vs. product returns), or even have the same intents but provide different levels of slot information (Ruane et al., 2018). Additionally, information is typically provided over multiple conversation turns, and each turn could involve multiple user intents and slots. Finally, the same set of tasks can be expressed using numerous possible utterances by users, depending on their choice of phrasing, order of sentences, use of colloquialisms, introducing digressions, etc. (Guichard et al., 2019). Hence, relying solely on distance based similarity of utterance embeddings would adversely impact performance.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16234–16240
December 6-10, 2023

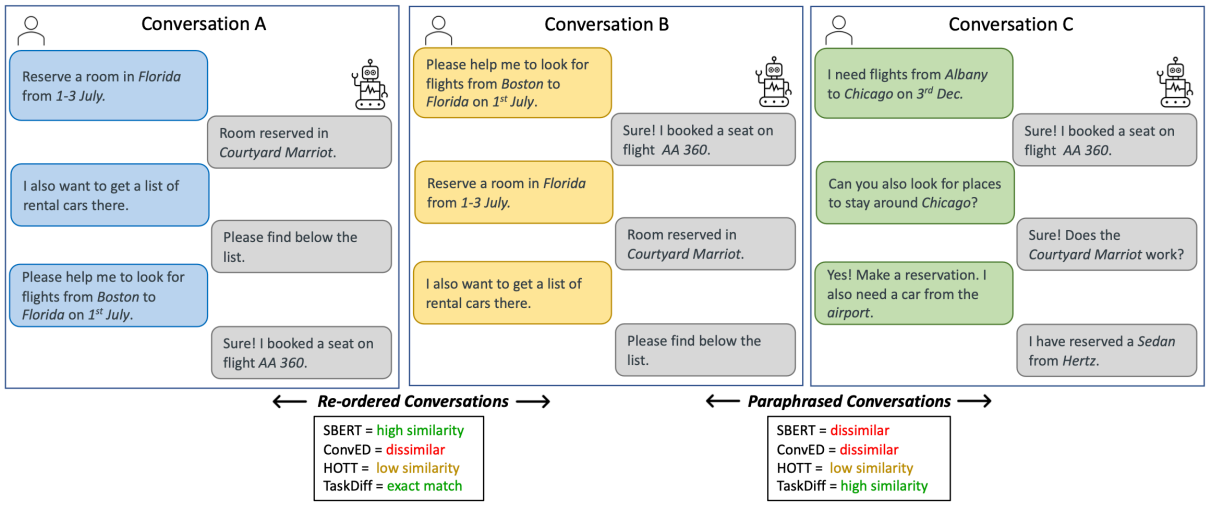

Figure 1: Demonstrating robustness of *TaskDiff* over prior approaches for multiple conversational scenarios.

In this work, we present *TaskDiff*, a novel similarity metric designed for task-oriented conversations to address the above challenges. Figure 1 shows multiple users having similar conversations about making bookings for a trip but with re-ordered tasks or paraphrased utterances with different slot values. It also shows that prior work is not robust to such differences that commonly occur in task-oriented conversations.

An ideal metric to measure conversational similarity should be able to identify that the overall goal of these conversations in Figure 1 is the same. *TaskDiff* represents the structure of conversations as distributions over the different task-oriented components and combines the geometry of the distributions with optimal transport to measure the similarity between conversations. Our approach is inspired by prior work in topic modelling (Kusner et al., 2015; Yurochkin et al., 2019) that have shown the effectiveness of comparing the structure of distributions, albeit for different settings. We evaluate *TaskDiff* on a benchmark task-oriented conversation dataset and demonstrate its effectiveness while presenting examples illustrating its improvement over existing approaches.

2 Task-Oriented Conversation Similarity

2.1 Definitions

A task-oriented conversational system supports a pre-defined set of user intents $\mathcal{I}$ and their corresponding slots or parameters $\mathcal{S}$. Each conversation C_i consists of a multi-turn sequence of utterances U_i between the user and the system or agent, a subset of active intents, and slot-value information provided by the user (i.e.) $I_i \subseteq \mathcal{I}$ and $S_i \subseteq \mathcal{S}$. Our objective is to compute the similarity between task-oriented conversations, given their components $K = \{U, \mathcal{I}, \mathcal{S}\}$ (i.e.) utterances, intents, and slot information.

2.2 Approach

TaskDiff measures similarity between task-oriented conversations as a function of the distance between their component-wise distributions. For each component $k \in K$, we represent its distribution over every conversation and compute similarity as the cumulative cost of transforming or transporting the component-wise distributions of one conversation to another.

Figure 2 shows an overview of *TaskDiff*. We first mask the values of the slots in every conversation with their corresponding '<slot name>' from the ontology, before using SBERT to generate conversational embeddings. The masking ensures that entities representing the slot values do not incorrectly bias or ambiguate the embeddings (Shi et al., 2018). For instance, the embedding similarity between the unrelated utterances - *"I want a ticket to the Big Apple"* and *"I want a ticket to the Apple conference"*, could be incorrectly influenced by the word 'Apple', but masking with their appropriate slot names (e.g., <arrival_city> and <product_name>), resolves this possibility. We denote Δ_U^l as the distribution of utterance embeddings of a single conversation.

We then compute probability distributions $\Delta_{\mathcal{I}}^n$, $\Delta_{\mathcal{S}}^m$ for each conversation over the set of intents $\mathcal{I}$

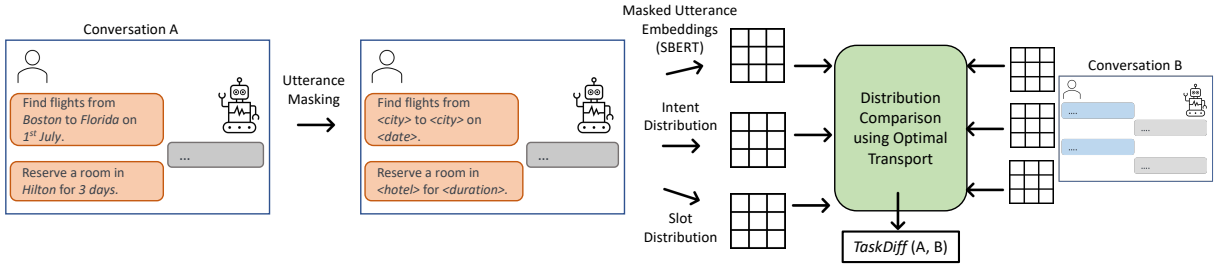

Figure 2: Overview of *TaskDiff* illustrating steps for masking utterances, generating distributions over different conversation components, and computing similarity using Optimal Transport cost between conversations' distributions.

and slots $\mathcal{S}$ as –

$$\Delta_K^n = \{p_i \in \mathbb{R}^{n+1} \mid \sum_{i=0}^{n} p_i = 1 \ , \ p_i \geq 0 \ \forall i \in |K|\}$$

where each p_i reflects the frequency of occurrence of intents and slots over the utterances. For example, $\Delta_{\mathcal{I}}^n$ for conversation C_i represents the probability of all n intents within C_i.

We then compute a separate cost matrix $\mathcal{M}_{i,j}$ for each component, that represents the cost to move between two points (i, j) in its distribution. We compute each entry using the Euclidean distance between the embeddings generated for each component. Intuitively, conversations with similar intents, slot information, and analogous language would reflect similar distributions, and hence a lower cost of transportation (i.e.) high similarity. However, any differences in their components would incur a larger cost, and hence reflect a lower similarity.

Given distributions $\alpha \in \Delta_k^a$, $\beta \in \Delta_k^b$, $\forall k \in K$ and the cost matrix $\mathcal{M}$, the 1-Wasserstein optimal transport distance (Vallender, 1974) between them is –

$$W_1(p, q) = \min_{\Gamma \in \mathbb{R}^{n \times m}} \sum\nolimits_{i,j} \mathcal{M}_{i,j} \Gamma_{i,j}$$

$$\text{subject to } \sum\nolimits_j \Gamma_{i,j} = \alpha_i \text{ and } \sum\nolimits_i \Gamma_{i,j} = \beta_j$$

where $\mathcal{M}_{i,j} = d(i, j)$ denotes the cost matrix and $d(.,.)$ denotes the distance between the distributions. We then define the similarity (*TaskDiff*) between two task-oriented conversations C_1 and C_2 as the weighted sum of the W_1 distances between their respective components –

$$\text{TaskDiff}(C_1, C_2) = \sum_{k=1}^{|K|} \gamma_k W_1(C_1^{\oplus}, C_2^{\oplus}) \quad (1)$$

where $C_i^{\oplus} = \{U_i, I_i, S_i\}$ represents the conversation's components K (i.e.) utterances, intents, and slots, and γ_k is a hyperparameter reflecting the influence of each component on the similarity.

3 Experimental Evaluation

3.1 Dataset

We use SGD (Rastogi et al., 2020), a benchmark dataset of multi-turn task-oriented conversations between users and agents spanning 20 domains (e.g., travel, dining). Its 20,000 conversations are annotated with active intents and slot information.

3.2 Baselines

We compare *TaskDiff* to three existing approaches:

1. **SBERT**: A state-of-the-art approach to measure similarity between conversational embeddings using cosine similarity (Reimers and Gurevych, 2019).

2. **Conversational Edit Distance (ConvED)**: A dialogue similarity metric that aligns utterances between conversations and computes the edit distance between their embeddings (Lavi et al., 2021).

3. **Hierarchical Optimal Transport (HOTT)**: A document similarity metric that by models topics using Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and subsequently uses the 1-Wasserstein distance on the topic and text embeddings (Yurochkin et al., 2019).

We conduct our experiments on an Intel Core i9 with 64GB of RAM. We implement *TaskDiff* in Python, leveraging the POT library (Flamary et al., 2021) for the 1-Wasserstein optimal transport distance. The choice of γ was set to 2, 1, and 1 for the intent, utterances and slots components, respectively after performing hyper-parameter search.

3.3 k-NN Classification

We evaluate the ability of the different approaches to accurately classify similar SGD conversations into the correct domains using k-NN. From Table

1, we observe that *TaskDiff* outperforms SBERT, HOTT and ConvED, demonstrating the importance of considering other conversational components for similarity beyond just utterances (i.e.) intents and slots, and the need for masking to avoid the adverse influence of entities. The utterance alignment coupled with use of edit distance in ConvED helps compared to SBERT, but requires annotations for alignment that may not always be available. We also see that HOTT returns the lowest accuracy, since LDA often picks topics outside the actual conversational intents due to its reliance on word-frequencies. This incorrectly skews the optimal transport distributions thereby impacting classification.

Approach	Accuracy
SBERT	0.78
HOTT	0.15
ConvED	0.86
TaskDiff	**0.95**

Table 1: Accuracy scores for k-NN classification

3.4 Conversational Clusters

We visualize the conversational clusters formed by the different approaches on SGD using k-means, setting k to 20 (i.e.) the number of domains and running 20 iterations. From Figure 3, we observe that *TaskDiff* results in the most well-formed and distinct clusters followed by SBERT, which has some cluster overlap and lower distinction. The clusters resulting from ConvED and HOTT show a significant amount of overlap, demonstrating their inability to distinguish between similar and dissimilar conversations.

3.5 Ablation Study

We perform an ablation study using 200 randomly selected dialogues, to highlight the influence of the different components in *TaskDiff* that enable its effectiveness over approaches like SBERT. As shown in Table 2, masking the slot names within the utterances results in a 14% improvement in accuracy over SBERT, since the embedding similarity is no longer influenced by incorrect biases or ambiguity from the slot values as described in Section 2.2.

Additionally, we see that the use of optimal transport (OT) based similarity on the utterances without the use of masks, suffers from the same drawbacks compared to when masks are introduced. Finally, the addition of intents and slots to the optimal transport (i.e.) *TaskDiff* results in a 26% improvement

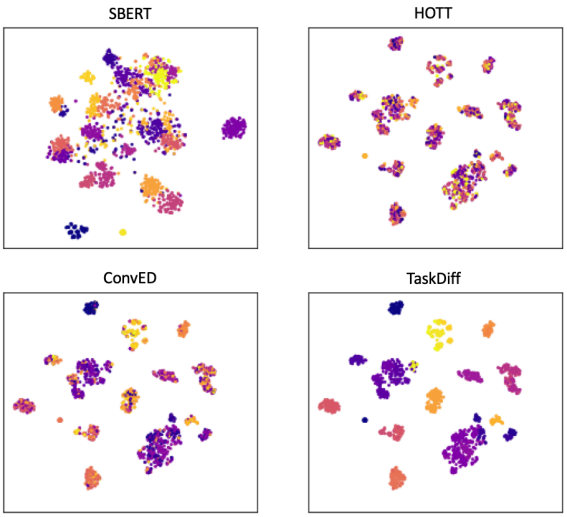

Figure 3: Conversations clustered using k-means and color coded by domain names.

in accuracy over SBERT, due to the additional information about the dialogues provided by these components, thereby highlighting their importance while measuring task-oriented conversation similarity.

Approach	Accuracy
SBERT	0.73
SBERT + Masking	0.83
SBERT + OT	0.68
SBERT + OT + Masking	0.85
TaskDiff	**0.92**

Table 2: Ablation study of *TaskDiff* with k-NN classification accuracy

3.6 Robustness to Reordering

We evaluate the robustness of the approaches for a common setting where users provide the same tasks in a different order within the conversation. We perturb the SGD dataset, wherein 30% of the utterances in each conversation are reordered, and compute their distance from the original for each approach. The average distance over all perturbed conversations in Table 3 shows that *TaskDiff* returns an exact match on these conversations, since representing conversations as distributions over its components (i.e., intents, slots, utterances), makes it agnostic and robust to such changes. The comparison approaches however, are not as robust, with ConvED performing poorly due to its reliance on alignments between utterances.

Approach	Avg. Distance
SBERT	0.005
HOTT	0.200
ConvED	4.150
TaskDiff	**0.000**

Table 3: Impact of conversational reordering

4 Related Work

Efforts across many natural language tasks including sentiment analysis (Poria et al., 2016), recommendation systems (Magara et al., 2018), and question answering (Sidorov et al., 2015), have relied on using distance-based similarity measures over text embeddings (Wang and Dong, 2020). Furthermore, recent work on dialogue similarity have also leveraged conversation structure, where Appel et al. (2018) consider the number of dialogue turns, words, and cycles and use cosine similarity. Similarly, Xu et al. (2019) cluster user-bot dialogues using different distance measures and Enayet and Sukthankar (2022) measure similarity of dialogue sequences using the Hamming distance.

The use of optimal transport over text distributions has shown promising results in document similarity (Solomon, 2018) resulting in popular metrics like the word mover's distance (WMD) (Kusner et al., 2015) and supervised WMD (Huang et al., 2016). Recently, Yurochkin et al. (2019) used optimal transport over topic models for documents, demonstrating a significant improvement in performance over traditional distance based measures. However, direct application of such approaches to task-oriented dialogues is challenging, due to the unique structure and different components of conversations, as shown in our results.

5 Conclusion

In this paper we present *TaskDiff*, a novel metric to measure the similarity between task-oriented conversations. It not only captures semantic similarity between the utterances but also utilizes dialog specific features like intents and slots to identify the overall objective of the conversations. We demonstrate that unlike existing metrics, taking advantage of these unique components is critical and results in significantly improved performance. As part of future work, we will investigate the inclusion of additional dialog features on open domain dialog datasets and the utilization of *TaskDiff* to improve the performance of various downstream conversational tasks.

6 Limitations

We demonstrate in this work that *TaskDiff* is a superior and more robust similarity metric compared to existing state-of-the-art approaches for task-oriented conversations. Given the use of optimal transport to compute similarity as a function of differences over the component distributions (intents, slots, and utterances), *TaskDiff* is reliant on being given an ontology for the intents and slots present across the conversations. However, this is a fair assumption to make for the domain of task-oriented conversations, and such ontologies are leveraged by real-world deployments such as Google DialogFlow, IBM Watson Assistant, Amazon Lex, etc.

References

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, et al. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Ana Paula Appel, Paulo Rodrigo Cavalin, Marisa Affonso Vasconcelos, and Claudio Santos Pinhanez. 2018. Combining textual content and structure to improve dialog similarity. *arXiv preprint arXiv:1802.07117*.

Sujoy Bag, Sri Krishna Kumar, and Manoj Kumar Tiwari. 2019. An efficient recommendation generation using relevant jaccard similarity. *Information Sciences*, 483:53–64.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Ayesha Enayet and Gita Sukthankar. 2022. An analysis of dialogue act sequence similarity across multiple domains. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3122–3130.

Rémi Flamary et al. 2021. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Jianfeng Gao, Chenyan Xiong, and Paul Bennett. 2020. Recent advances in conversational information retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2424.

Jonathan Guichard, Elayne Ruane, Ross Smith, Dan Bean, and Anthony Ventresque. 2019. Assessing the robustness of conversational agents using paraphrases. In *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, pages 55–62. IEEE.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. 2016. Supervised word mover's distance. *Advances in neural information processing systems*, 29.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Ofer Lavi, Ella Rabinovich, Segev Shlomov, David Boaz, Inbal Ronen, and Ateret Anaby Tavor. 2021. We've had this conversation before: A novel approach to measuring dialog similarity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1169–1177.

Shuyang Li, Bodhisattwa Prasad Majumder, and Julian McAuley. 2022. Self-supervised bot play for transcript-free conversational recommendation with rationales. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 327–337.

Maake Benard Magara, Sunday O Ojo, and Tranos Zuva. 2018. A comparative analysis of text similarity measures and algorithms in research paper recommender systems. In *2018 conference on information communications technology and society (ICTAS)*, pages 1–5. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Federica Bisio. 2016. Sentic lda: Improving on lda with semantic similarity for aspect-based sentiment analysis. In *2016 international joint conference on neural networks (IJCNN)*, pages 4465–4473. IEEE.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Elayne Ruane, Théo Faure, Ross Smith, Dan Bean, Julie Carson-Berndsen, and Anthony Ventresque. 2018. Botest: a framework to test the quality of conversational agents using divergent input examples. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, pages 1–2.

Yong Shi, Yuanchun Zheng, Kun Guo, Wei Li, and Luyao Zhu. 2018. Word similarity fails in multiple sense word embedding. In *International Conference on Computational Science*, pages 489–498. Springer.

Grigori Sidorov, Helena Gómez-Adorno, Ilia Markov, David Pinto, and Nahun Loya. 2015. Computing text similarity using tree edit distance. In *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*, pages 1–4. IEEE.

Justin Solomon. 2018. Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

SS Vallender. 1974. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

Jiapeng Wang and Yihong Dong. 2020. Measurement of text similarity: a survey. *Information*, 11(9):421.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Luxun Xu, Vagelis Hristidis, and Nhat XT Le. 2019. Clustering-based summarization of transactional chatbot logs. In *2019 IEEE International Conference on Humanized Computing and Communication (HCC)*, pages 60–67. IEEE.

Avi Yaeli, Segev Shlomov, Alon Oved, Sergey Zeltyn, and Nir Mashkif. 2022. Recommending next best skill in conversational robotic process automation. In *International Conference on Business Process Management*, pages 215–230. Springer.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. 2019. Hierarchical optimal transport for document representation. *Advances in Neural Information Processing Systems*, 32.