

Remembering the Flow: Retrieval with Dynamic Turn Representations for Conversational Memory

Juan Manuel Fernández^{1,2}, Marcelo Errecalde³, and Sergio Burdisso⁴

¹ LICDIA, National University of Luján, Luján, Argentina

² Buenos Aires Institute of Technology (ITBA), Buenos Aires, Argentina

³ LIDIC, National University of San Luis, San Luis, Argentina

⁴ Idiap Research Institute, Martigny, Switzerland

jmfernandez@{unlu,itba}.edu.ar, merreca@unsl.edu.ar, sburdisso@idiap.ch

Abstract. In task-oriented dialogue, conversational turns are represented as dense embedding vectors that drive the agent’s decisions. However, dialogues are sequential: the meaning of a turn depends on its accumulated history, while turns are typically encoded in isolation. We study how much of that history a turn representation should carry, comparing static embeddings, a normalized cumulative state, and a calibrated exponential moving average (EMA). Usefulness is measured by retrieving past situations and scoring them with an LLM judge (MSS@10) on held-out, cross-dialogue queries. Over a conversation-specific space (Dialog2Flow), even a simple average of past turns outperforms isolated turns, and EMA ($\alpha = 0.6$) improves further and generalizes to unseen queries; over general-purpose encoders the same accumulation does not help. The value of remembering thus depends on the base geometry, not accumulation itself, and encoding history within the turn representation is an effective route to conversational memory when that space is conversation-aware.

Keywords: Task-oriented dialogue, Conversational memory, Dynamic turn representations, Dialogue embeddings, Dialog2Flow.

1 Introduction

Task-oriented dialogue systems help a user accomplish a concrete goal—booking a hotel, requesting a ride, querying a service—by interpreting each turn, maintaining a representation of the conversational state, and choosing an action at every step [29,5]. Turns are commonly represented as sentence embeddings: dense vectors from pretrained models that place semantically related utterances nearby [20,6].

Dialogue, however, unfolds as a sequence, and many turns are uninterpretable in isolation: elliptical answers, references, and confirmations resolve only against the preceding turns. This is precisely why task-oriented systems track a dialogue or belief state across turns [29,27], and why a growing body of work induces

latent representations of the conversational context rather than of the isolated utterance [26,16,9]. An embedding computed turn by turn ignores this trajectory, which motivates *dynamic* representations that fold the accumulated history into each turn.

Turn representations can also serve as queries into a memory of past interactions. Approximate nearest neighbor (ANN) search retrieves the most similar entries from very large collections at low latency [11,15,10], a natural substrate for conversational memory. By this we do not mean what the term usually denotes for large language models—compressing or summarizing the ongoing conversation [28,17]—but an episodic store: the agent retrieves past situations similar to its current one, typically from other conversations, as additional context to inform—or verify—its next action [18,31]. The value of such a memory depends on whether the representation places *functionally* similar situations nearby.

Concretely, we build a retrieval setting over a collection of about one million task-oriented turns, index it with ANN structures, and use the three turn representations—static, normalized cumulative, and EMA-calibrated—as query encoders. We ask how much conversational history a turn representation should carry, and assess the retrieved neighborhoods with an LLM-assisted semantic-functional metric (MSS@10) under a cross-dialogue protocol that excludes the query’s own conversation from the memory.⁵

2 Background

Task-oriented dialogue has traditionally been framed as sequential interaction, with the agent interpreting the conversational state and selecting actions toward a goal; classical approaches modeled this through explicit dialogue and belief states, often as (partially observable) decision processes [29,27,5]. Neural models then replaced hand-crafted symbols with distributed representations: embeddings encode words, sentences, or turns as dense vectors [2,20], and Transformer architectures extend them toward contextualized representations [24,6]. Applied independently to each turn, however, such embeddings capture the local content of an utterance rather than the position of the turn within the conversational trajectory. This gap has motivated work that induces latent dialogue states, actions, or structures from data—summarizing the history in a distributed state [26] or inducing latent structure with little supervision [16,9,13]—the line to which the dynamic turn representations studied here belong.

Representing turns well, however, is only half the problem: to act as a memory, these vectors must be searchable at scale. Approximate nearest neighbor (ANN) search meets this need through space partitioning, navigable graphs, and quantization, trading accuracy for speed and memory [22,15,10,11]. The combination of embeddings and ANN has scaled semantic search, question answering, retrieval augmentation, and response selection [7,21,19]. In LLM-based dialogue systems the same machinery underlies retrieval-augmented generation

⁵ Code, index parameters, judge prompts, and full result tables are available at <https://github.com/jumafernandez/remembering-the-flow>.

(RAG) [12], where retrieved evidence conditions the next response; yet the retrieved units are documents or candidate responses—conversational *states* or situations remain comparatively unexplored.

Two elements are specific to our study. First, Dialog2Flow pretrains turn embeddings guided by communicative actions, organizing task-oriented utterances into a latent space of dialogue functions [4]; it provides the conversation-aware space over which we study accumulation. Second, because geometric fidelity to exact neighbors does not reveal whether retrieved turns are *functionally* plausible, we adopt LLM-assisted evaluation, following work that uses language models as judges of semantic quality [14,30].

These threads—dynamic conversational representations, scalable approximate retrieval, and functional evaluation—rarely meet. We integrate them: a dialogue becomes a memory of turn representations, we ask how much history each representation should carry, and we judge the retrieved neighborhoods not by geometric recall but by their plausibility as conversational memory. Whether accumulation helps, and how it should be calibrated, then becomes an empirical question about the interplay between the base embedding space and the update policy.

3 Methodology

This section describes the data, the turn representations we compare, and the retrieval-based, semantic-functional protocol used to evaluate them.

3.1 Dataset

We build an experimental collection from Dialog2Flow [4], which unifies several task-oriented dialogue corpora under a common format. We retain 101,021 complete dialogues comprising 1,000,023 turns, preserving dialogue integrity: turns are added until the one-million-vector threshold is exceeded, always after finishing the dialogue in progress. Each turn is encoded from its utterance text; dialogue and turn identifiers preserve the conversational order.

3.2 Turn representations

Each encoder maps a turn to a vector in its own embedding space, so comparing encoders amounts to comparing the geometries those spaces induce. We use three: two general-purpose Sentence-BERT checkpoints [20], all-MiniLM-L6-v2 [25] and all-mpnet-base-v2 [23], and one conversation-specific checkpoint, dialog2flow-joint-bert-base [4], which produce 384-, 768-, and 768-dimensional vectors, respectively. Evaluating across them lets us ask whether the value of accumulation depends on the base geometry.

Formally, a dialogue is a sequence of turns $u_1, \dots, u_T$, and an encoder f_{enc} maps each turn to a vector $e_t = f_{\text{enc}}(u_t)$. The question we study is what a turn’s representation h_t should contain: only e_t , or also the history $e_1, \dots, e_{t-1}$. Figure 1

contrasts the representations we compare over each of these spaces. In the *static* case (Fig. 1, top), each turn is encoded independently, $e_t = f_{\text{enc}}(u_t)$, so its vector ignores the surrounding conversation. The *dynamic* representations (Fig. 1, bottom) instead fold the accumulated history into each turn: the *normalized cumulative* state adds the running history, $h_t = \text{LayerNorm}(h_{t-1} + e_t)$ [1] (reset at each dialogue, $h_0 = 0$), and the *EMA* representation [3,8] generalizes it by weighting history against the current turn, $h_t^{(\alpha)} = \text{LayerNorm}((1-\alpha)h_{t-1} + \alpha e_t)$ with $\alpha \in \{0.1, \dots, 0.9\}$, whose effect we examine in Section 4. All three are post-hoc policies over frozen embeddings, not trained models, and the EMA family is evaluated over the Dialog2Flow space. The static representation is our baseline.

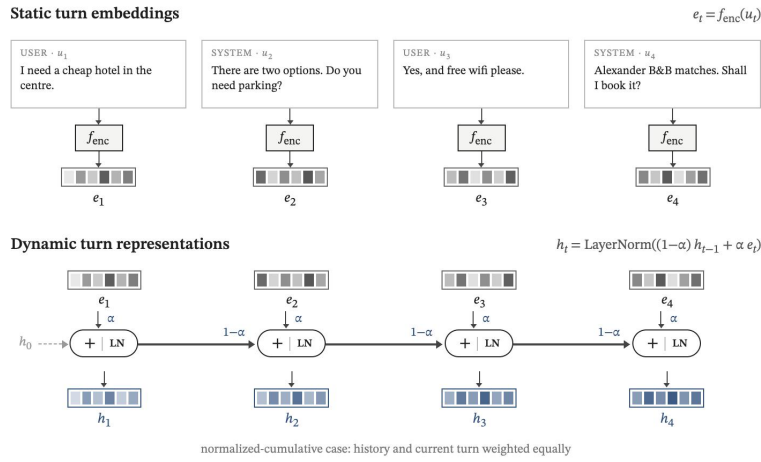


Fig. 1. Turn representations over a running example: *Static* (top), each turn encoded independently; *Dynamic* (bottom), folding accumulated history into each turn.

3.3 Retrieval and evaluation

For retrieval, the collection of turn vectors is split into a query set Q and an index set V with $Q \cap V = \emptyset$ ($|Q| = 10,000$, $|V| = 990,023$), so a query never retrieves its own vector; the split is fixed across all representations. Vectors are L_2 -normalized—so inner product ranks by cosine similarity—and indexed with three FAISS structures [11] spanning the main ANN families—inverted lists (IVF), a navigable graph (HNSW [15]), and product quantization (IVFPQ [10])—whose parameters are tuned to high-recall operating points; exact values are in the public repository. Retrieval quality is reported as Recall@ k against exact search (FlatL2), alongside queries per second and memory, summarizing the accuracy–speed–memory trade-off.

Geometric recall, however, does not reveal whether the retrieved turns are *functionally* plausible as conversational memory. We therefore add an LLM-

assisted, semantic-functional evaluation: for each query q , a judge (gpt-4.1-mini, temperature 0) rates every neighbor v_i in the retrieved top-10 for semantic and functional similarity to the query situation, returning an overall *similarity score* $\text{score}(q, v_i) \in \{1, \dots, 5\}$ [14,30]. We summarize a query by its mean top-10 similarity score (SS@10) and a configuration by the mean over the query set Q (MSS@10),

$$\text{SS@10}(q) = \frac{1}{10} \sum_{i=1}^{10} \text{score}(q, v_i), \quad \text{MSS@10} = \frac{1}{|Q|} \sum_{q \in Q} \text{SS@10}(q), \quad (1)$$

with $|Q| = 100$ judged queries in the main evaluation, extended to $|Q| = 500$ (those 100 plus 400 held-out queries) under the cross-dialogue protocol.

Two safeguards make the comparison trustworthy. First, because the split still allows retrieving *other* turns of the query’s own dialogue, a query-time filter discards same-dialogue candidates, making retrieval strictly cross-dialogue. Second, the 400 held-out queries are disjoint from those used to select the EMA weight α . Variants are compared per query with the two-sided Wilcoxon signed-rank test on paired SS@10 differences, reporting also the fraction of improving queries.

4 Experiments and results

The experimental analysis begins with a first stage in which the performance of the ANN indexes over static representations is evaluated. For the three considered spaces, IVF, HNSW, and IVFPQ indexes were built, using FlatL2 as the exact reference for the computation of recall (Figure 2).

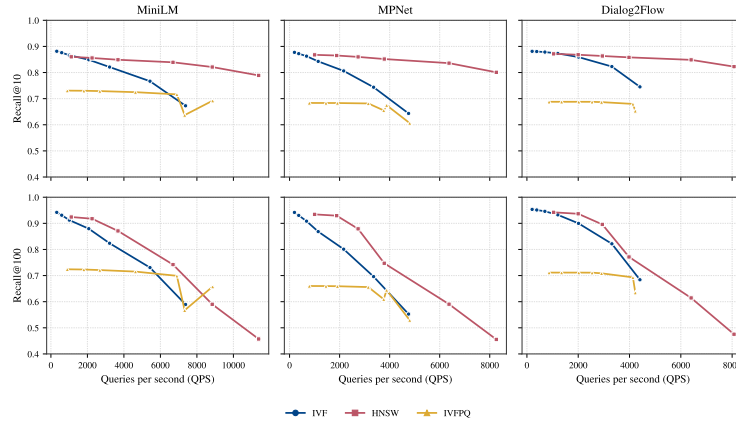


Fig. 2. Approximate retrieval over static embeddings: Recall@10 and Recall@100 versus QPS (FlatL2 reference), for the three spaces and IVF/HNSW/IVFPQ.

IVF reaches Recall@10 near 0.88 and Recall@100 near 0.94, HNSW sustains comparable recall at more than 1000 queries per second, and IVFPQ compresses the index roughly 30–60× (to about 50 MB) at a moderate recall cost. Notably, the conversation-aware Dialog2Flow space indexes as efficiently as the general-purpose encoders. Approximate retrieval is therefore a controlled substrate: the differences reported below are properties of the representation, not of the index.

A faithful index, however, is not yet a useful memory. Recall guarantees that each space returns its true nearest neighbors, but not that those neighbors are the *right* ones to recall: what an agent needs from memory is a past turn that is *functionally* analogous to the present one—a comparable moment in another dialogue, where the earlier decision can inform the current one. Two representations may be indexed equally well and still differ sharply on this, and geometric recall, which only measures fidelity to exact search, cannot distinguish them.

To expose that difference we turn to the semantic-functional judgment of Section 3: for each query we take its top-10 cross-dialogue neighbors and let an LLM judge rate how closely each matches the query situation, aggregated per configuration as MSS@10 (Table 1). This lets us pose the central question of the study—does folding conversational history into the turn representation retrieve *more useful* memories, and does the answer depend on the underlying embedding geometry? What we find is a clean dissociation.

Table 1. Semantic-functional evaluation: MSS@10 (mean $\pm$ SD over $n=100$ queries, 1–5 scale) for the static (baseline) and cumulative variants in each space and index.

Space	Variant	IVF	HNSW	IVFPQ
MiniLM	Static	3.40 $\pm$ 0.76	3.42 $\pm$ 0.74	3.35 $\pm$ 0.80
MiniLM	Cumulative	3.32 $\pm$ 0.75	3.33 $\pm$ 0.76	3.23 $\pm$ 0.77
MPNet	Static	3.38 $\pm$ 0.77	3.36 $\pm$ 0.80	3.37 $\pm$ 0.82
MPNet	Cumulative	3.31 $\pm$ 0.78	3.31 $\pm$ 0.76	3.20 $\pm$ 0.74
Dialog2Flow	Static	3.29 $\pm$ 0.80	3.29 $\pm$ 0.77	3.30 $\pm$ 0.74
Dialog2Flow	Cumulative	3.66 $\pm$ 0.59	3.66 $\pm$ 0.58	3.65 $\pm$ 0.63

Accumulation helps *only* in the conversation-aware space: over Dialog2Flow the cumulative representation lifts MSS@10 from the static baseline (about 3.29) to about 3.66 across all three indexes—the highest scores in the table—while over the general-purpose encoders MiniLM and MPNet the same operation is flat to slightly negative.

Two things would make this dissociation convincing: that the Dialog2Flow gain is broad across queries rather than the work of a few, and that it tracks the underlying geometry—that accumulation should help precisely where it leaves the neighborhood structure intact. Table 2 evaluates both, setting the per-query

win rate of cumulative over static beside how far accumulation moves each space, measured by cosine similarity and neighbor overlap.

Table 2. Geometry (cosine, Overlap@100 of static vs. cumulative embeddings) and per-query win rate of cumulative over static, by space and index.

Space	Geometry		Wins over static (%)		
	cos	Overlap@100	IVF	HNSW	IVFPQ
MiniLM	0.412	0.035	47	48	44
MPNet	0.405	0.032	48	48	41
Dialog2Flow	0.727	0.152	68	65	68

Both hold. Over Dialog2Flow, accumulation barely disturbs the neighborhoods (cosine 0.73, Overlap@100 0.15) and the cumulative variant wins 65–68% of paired queries (Wilcoxon $p < 2 \times 10^{-3}$); over MiniLM and MPNet it overturns the neighborhood structure (Overlap@100 ≈ 0.03) and the win rate falls to parity (41–48%, none significant). Geometry and function move together: where remembering preserves the local structure it also proves beneficial query by query—the coupling the MSS@10 gap already hinted at.

Figure 3 examines Dialog2Flow to show what that reorganization looks like, and two observations emerge.

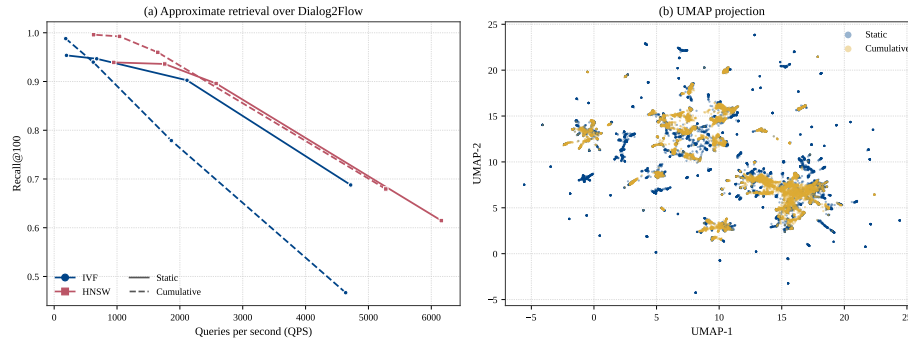


Fig. 3. Exploratory view of the Dialog2Flow space. (a) Recall@100 versus QPS for IVF and HNSW (against exact search); (b) UMAP projection of static and cumulative turn embeddings.

First, the reshaping carries no retrieval penalty: in panel (a), the Recall@100–QPS curves show the cumulative representation staying as searchable as the static one—and somewhat more so at high-recall operating points—so the functional gain costs nothing in approximate-search efficiency. Second, panel (b)

makes the restructuring visible: the space genuinely rearranges under accumulation, consistent with the modest neighbor overlap in Table 2. Crucially, this is the *constructive* kind of reorganization: unlike in the general-purpose spaces, the re-arrangement coincides with the semantic–functional gain rather than obscuring it.

So far, accumulation has been uniform: every past turn contributes equally to the running representation. But that is only one point on a spectrum—how heavily a turn should lean on its accumulated history, rather than on its own content, is itself a tunable quantity. The EMA update exposes it through the weight α , which fixes how much of a representation comes from the current turn and how much from its history: $\alpha=0.9$ makes it 90% current turn and 10% history, while $\alpha=0.5$ weights both equally and recovers the uniform cumulative baseline. We sweep α over Dialog2Flow—the only space where accumulation proved beneficial—to ask whether a sharper balance exists (Fig. 4).

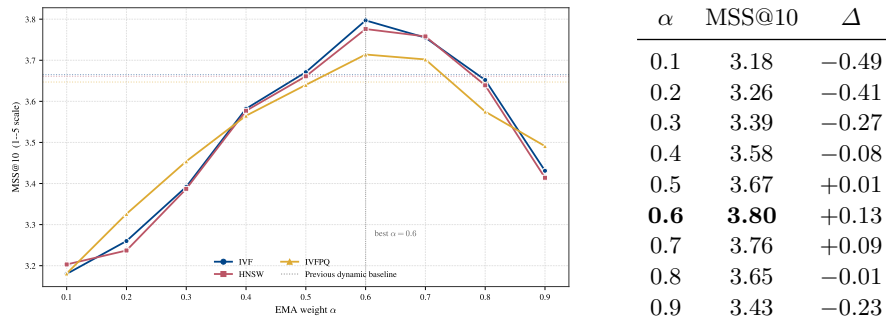


Fig. 4. EMA weight sweep over Dialog2Flow: MSS@10 vs. α (left) and its gain Δ over the cumulative baseline (right).

It does, though by a thin margin. MSS@10 peaks at $\alpha=0.6$ and declines on either side—small α even trails the uniform baseline. At the peak, the gain over that baseline is +0.13, +0.12, and +0.07 MSS@10 for IVF, HNSW, and IVFPQ, a broad win rather than an outlier (52/17/31 wins/ties/losses on IVF). At $n=100$, though, the improvement is only borderline ($p \approx 0.06$); a stricter, larger-scale cross-dialogue test resolves it (Table 3).

That test addresses a concrete concern: the vector-level split can still return other turns from the query’s own dialogue. Re-evaluating all three representations on 500 queries (the original 100 plus 400 held-out), the ordering is preserved—static ≈ 3.35 , cumulative ≈ 3.72 , EMA ≈ 3.79 —and the paired tests confirm both effects: the static-to-cumulative gain survives with a wide margin, and the calibrated EMA now *significantly* outperforms the uniform cumulative baseline in every index, held-out queries included—the out-of-sample result that resolves the borderline $n = 100$ comparison.

Table 3. Cross-dialogue MSS@10 over Dialog2Flow, paired Wilcoxon tests. Δ : mean per-query difference; p : two-sided. *Full*: 500 queries; *Held-out*: 400.

Contrast		IVF		HNSW		IVFPQ	
		Δ	p	Δ	p	Δ	p
Cumulative – Static	Full	+0.37	9×10^{-25}	+0.38	2×10^{-25}	+0.35	2×10^{-23}
	Held-out	+0.36	1×10^{-19}	+0.36	1×10^{-19}	+0.35	6×10^{-18}
EMA – Cumulative	Full	+0.08	2×10^{-5}	+0.07	4×10^{-4}	+0.07	5×10^{-4}
	Held-out	+0.08	7×10^{-5}	+0.07	8×10^{-4}	+0.08	1×10^{-3}
EMA – Static	Full	+0.45	2×10^{-34}	+0.44	1×10^{-34}	+0.42	1×10^{-32}
	Held-out	+0.44	5×10^{-27}	+0.43	2×10^{-26}	+0.43	3×10^{-26}

A near-duplicate control excludes one alternative reading. Task-oriented corpora reuse near-identical turns across conversations, and a representation that merely retrieved those copies would score well without capturing anything functional. We measure how often each variant’s top neighbors are such cross-dialogue near-duplicates (Table 4).

Table 4. Cross-dialogue near-duplicates over Dialog2Flow: share of top-10 and rank-1 neighbors that are near-identical turns from *other* dialogues.

Variant	Near-dup@10	Rank-1 near-dup
Static	16.4%	43.2%
Dynamic	3.8%	20.8%
EMA ($\alpha=0.6$)	3.8%	20.9%

The static representation is responsible for this: near-identical turns constitute 16.4% of its top-10 neighbors and are the nearest neighbor for 43% of queries, against about 4% and 21% for the dynamic variants. Because these duplicates draw markedly higher judge scores (4.18 versus 3.57), they *inflate* the static MSS@10—so the static-to-cumulative gap is, if anything, conservative, while the EMA-over-cumulative contrast, sitting at the same low duplicate rate, is untouched.

Together, these controls make the dynamic advantage difficult to dismiss: it is significant in every index, it holds on queries withheld from tuning, and it is not manufactured by near-duplicate retrieval. The gain from encoding history is thus a property of the representation and its geometry, not of the evaluation—which is exactly what a memory of past conversations should require.

5 Conclusions and future work

Our question—how much of its accumulated history a turn representation should carry—depends less on the accumulation rule than on the space it operates in. Folding history into the turn helps retrieve functionally relevant past situations over a conversation-aware geometry—where a calibrated moving average helps a little more—but not over general-purpose encoders. Remembering is beneficial, in other words, only where the base space is already conversation-aware.

Two considerations bound the scope of these claims. First, our evidence comes from a single conversation-aware space, Dialog2Flow; the finding that accumulation helps over conversation-oriented representations is therefore established for Dialog2Flow specifically, not for the whole class of such spaces, and confirming it as a class-level property would require evaluating other conversation-aware embeddings. Second, functional plausibility is assessed by an LLM judge. Rather than a limitation, this is what makes it feasible to scale the number of queries and checks—paired comparisons, the cross-dialogue protocol, and held-out replications—to a volume out of reach for human annotation. The systematic scoring is thus delegated to the model; human oversight was qualitative and non-systematic, aimed at checking the coherence of the scores rather than constituting a formal annotation study.

The representations studied here are hand-set, post-hoc policies over frozen embeddings—a conservative, no-learning floor. The clearest next step is to *learn* the update instead of fixing it: a model over turn embeddings, trained to organize situations by communicative function, of which the cumulative and EMA policies are special cases. Further out, the fact that memory quality tracks conversational geometry suggests a longer-term direction—shaping representations whose neighborhoods correspond to shared dialogue *states*, inducing the transition structure of a task from them, and using that structure to constrain what an agent does next. Retrieval is the grounded first step: a memory can only guide behavior once its neighbors are the right situations to recall.

Acknowledgements. This research was conducted as part of a project supported by the IPAC initiative (High Performance Computing Argentina), under a PCI-146 (Intensive Computing Projects) call, a national program for allocating high-performance computing resources. The authors also acknowledge the National High Performance Computing System (SNCAD), the Clementina XXI infrastructure, SICyT, and its technical staff for their support in the execution of the experiments. Finally, the authors thank Nora Reyes and Edgar Chávez for their expertise and collaboration on topics related to high-dimensional proximity search.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)

2. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. *Journal of Machine Learning Research* 3, 1137–1155 (2003)
3. Brown, R.G.: *Smoothing, Forecasting and Prediction of Discrete Time Series*. Prentice-Hall, Englewood Cliffs, NJ (1963)
4. Burdisso, S., Madikeri, S., Motlicek, P.: Dialog2Flow: Pre-training soft-contrastive action-driven sentence embeddings for automatic dialog flow extraction. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics (2024)
5. Chen, H., Liu, X., Yin, D., Tang, J.: A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19(2), 25–35 (2017)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*. pp. 4171–4186 (2019)
7. Eric, M., Manning, C.D.: Key-value retrieval networks for task-oriented dialogue. In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. pp. 37–49 (2017)
8. Gardner, E.S.: Exponential smoothing: The state of the art—part II. *International Journal of Forecasting* 22(4), 637–666 (2006)
9. Hudeček, V., Dušek, O.: Learning interpretable latent dialogue actions with less supervision. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*. pp. 327–339 (2022)
10. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(1), 117–128 (2011)
11. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* 7(3), 535–547 (2021)
12. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *Advances in Neural Information Processing Systems*. vol. 33, pp. 9459–9474 (2020)
13. Liu, H., Cai, Y., Lin, Z., Ou, Z., Huang, Y., Feng, J.: Variational latent-state gpt for semi-supervised task-oriented dialog systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31, 1777–1792 (2023)
14. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023)
15. Malkov, Y.A., Yashunin, D.A.: Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(4), 824–836 (2020)
16. Min, Q., Qin, L., Teng, Z., Liu, X., Zhang, Y.: Dialogue state induction using neural latent variable models. *arXiv preprint arXiv:2008.05666* (2020)
17. Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S.G., Stoica, I., Gonzalez, J.E.: MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560* (2023)
18. Park, J.S., O’Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)* (2023)

19. Raposo, G., Ribeiro, R., Martins de Matos, D.: Prompting, retrieval and training: An exploration of different approaches for task-oriented dialogue generation. In: Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 443–458 (2023)
20. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. pp. 3982–3992 (2019)
21. Sahoo, S.R., Gupta, A., Singh, M.: Retrieval augmented generative task-oriented dialogue systems. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. pp. 460–470 (2022)
22. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1470–1477 (2003)
23. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: MPNet: Masked and permuted pre-training for language understanding. In: Advances in Neural Information Processing Systems. vol. 33, pp. 16857–16867 (2020)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
25. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Advances in Neural Information Processing Systems. vol. 33, pp. 5776–5788 (2020)
26. Wen, H., Liu, Y., Che, W., Qin, L., Liu, T.: Sequence-to-sequence learning for task-oriented dialogue with dialogue state representation. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3781–3792 (2018)
27. Williams, J.D., Young, S.: Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language* 27(2), 393–422 (2013)
28. Xu, J., Szlam, A., Weston, J.: Beyond goldfish memory: Long-term open-domain conversation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. pp. 5180–5197 (2022)
29. Young, S., Gašić, M., Thomson, B., Williams, J.D.: Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE* 101(5), 1160–1179 (2013)
30. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685* (2023)
31. Zhong, W., Guo, L., Gao, Q., Ye, H., Wang, Y.: MemoryBank: Enhancing large language models with long-term memory. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19724–19731 (2024)