# PENTATRON: PErsonalized coNText-Aware Transformer for Retrieval-based cOnversational uNderstanding

**Niranjan Uma Naresh**[*]    **Ziyan Jiang**[*]    **Ankit**[*]
**Sungjin Lee**    **Jie Hao**    **Xing Fan**    **Chenlei Guo**
Amazon
{niumanar, ziyjiang, ankitvys, sungjinl, jieha, fanxing, guochenl}@amazon.com

## Abstract

Conversational understanding is an integral part of modern intelligent devices. In a large fraction of the global traffic from people using smart digital assistants, frictions in dialogues may be attributed to incorrect understanding of the entities in a user's query due to factors including ambiguous mentions, mispronunciation, background noise and faulty on-device signal processing. Such errors are compounded by two common deficiencies from intelligent devices namely, (1) the device not being tailored to individual users, and (2) the device responses being unaware of the context in the conversation session . Viewing this problem via the lens of retrieval-based search engines, we build and evaluate a scalable entity correction system, PENTATRON. The system leverages a parametric transformer-based language model to learn patterns from in-session user-device interactions coupled with a non-parametric personalized entity index to compute the correct query, which aids downstream components in reasoning about the best response. In addition to establishing baselines and demonstrating the value of personalized and context-aware systems, we use multitasking to learn the domain of the correct entity. We also investigate the utility of language model prompts. Through extensive experiments, we show a significant upward movement of the key metric (Exact Match) by up to 500.97% (relative to the baseline).

## 1 Introduction

Intelligent devices are ubiquitous in the modern computing. The scientific modules that drive these devices involve conversational understanding, ambient computing, natural language reasoning and self-learning (Thoppilan et al., 2022; Sarikaya, 2022; Pinhanez et al., 2021; Liu et al., 2021). A user's interaction with a device, however, is susceptible to errors arising from a myriad of sources including wrong pronunciation, inaccuracies in

the subject mentions in a sentence, environmental noise, hardware and software error (Kim et al., 2020). Correct interpretations of user queries, especially entities, is central to delivering the best user experience. Two important factors that contribute strongly to high-precision entity recognition are (1) personalization, ie, learning users' unique patterns, and (2) contextualization, ie, deriving cues from the information in a user-device interaction session. In this paper, we design and evaluate an entity correction system, PENTATRON, with both personalization and contextualization baked into its architecture.
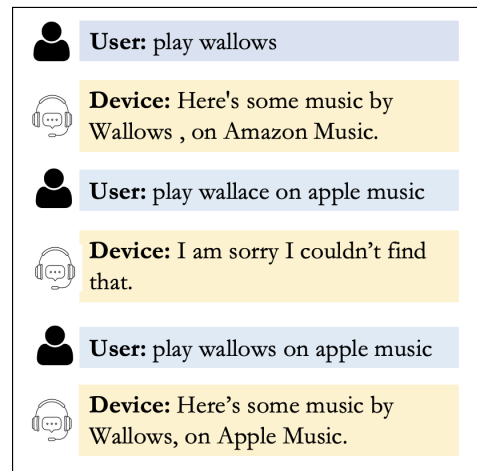


Figure 1: (Above) One multi-turn dialogue session with defective source query which contains one erroneous entity 'wallace' and its successful rephrase with correct entity 'wallows'. (Below) Concatenation of queries and responses using special tokens to form a single sequence as encoder input.

---

[*]Equal contribution.

## 1.1 Motivation

In Figure 1, we illustrate a real-world case as to why personalization and contextualization are very important, especially due to the specificity in highly entity-centric domains such as music. In this case, masking the very last device response, we observe that there is valuable information scattered across the user's requests in the session yet, the device delivers sub-par experience by responding defectively multiple times before finally getting the user's intent right.

## 1.2 Notation and Preliminaries

**Definition 1.** *Let integer $\gamma$ satisfy $1 \leq \gamma < \infty$. A natural language (NL) hypothesis is a mapping, $h : Q \rightarrow D \times I \times [E]^\gamma$, where $Q$ refers to the query space, $D$ refers to the domain space, $I$ refers to the intent space and $E$ refers to the entity space. The entity space, $E := E_T \times E_V$, may further be decomposed into the entity type space $E_T$ and the entity value space $E_V$. All spaces are defined over Unicode strings.*

As an example, given a query string $q =$"*play the real slim shady*", the corresponding NL hypothesis is $h(q) =$(*Music, PlayMusicIntent, [(SongName, the real slim shady)]*) where the domain is *Music*, the intent is *PlayMusicIntent*, and the entity value is *the real slim shady* with *SongName* entity type.

**Definition 2.** *Building on Definition 1, our system, PENTATRON, may be formalized as $\Phi : (C, Q) \rightarrow E_V$ where $C$ is the user space (anonymized using a hash function, for privacy, in practice).*

In a nutshell, given an input query $q$ (with or without dialogue context), our system essentially solves the optimization problem,

$$\min_\theta \mathbb{E}_{(c,q,e) \sim \mathcal{D}} \left[ \ell \left( \Phi_\theta(c, q), e \right) \right] \quad (1)$$

where $\mathcal{D}$ is supported on $C \times Q \times E_V$.

## 1.3 Our Contributions and Preview of Results

On the system design front, we build a retrieval-based pipeline. Our model backbone is inspired by attention-based (Vaswani et al., 2017) transformer encoders (Devlin et al., 2018). We achieve personalization via a non-parametric index which is essentially a key-value pair look-up table with the keys representing users and values representing the entity lists derived from historical data aggregation. With respect to experimental results, we
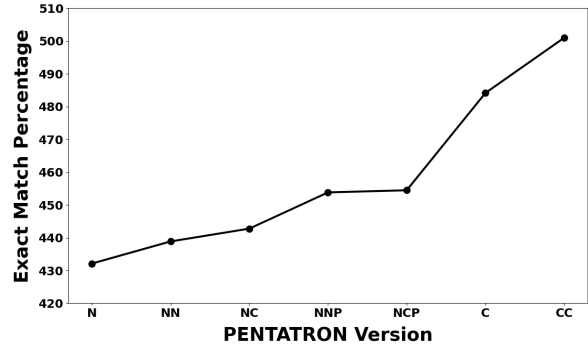


Figure 2: Preview of the system performance which shows consistent significant improvement in going from a purely personalized system (N) to a fully contextual personalized system (CC). Further details are available in Table 1.

conduct extensive studies on seven different versions of PENTATRON, involving ablations with prompts, multi-tasking and non-contextual training data, and show consistent improvements in Exact Match (EM) of up to 500.97% (relative to the baseline) as captured by the preview of results in Figure 2.

## 2 Background and Related Work

### 2.1 Query Rewriting

Query Rewriting (QR) in dialogue systems aims to reduce frictions by reformulating the automatic speech recognition component's interpretation of users' queries. Initial efforts (Dehghani et al., 2017; Su et al., 2019) treat QR as a text generation problem.

Some recent studies (Chen et al., 2020; Yuan et al., 2021; Fan et al., 2021; Cho et al., 2021) are based on neural retrieval systems. In the retrieval-based systems, the rewrite candidate pool is aggregated from users' habitual or historical queries so that the rewrite quality can be tightly controlled. Compared to generation-based systems, retrieval-based systems may sacrifice flexibility and diversity of the rewrites, but in the meanwhile provide more stability which is more important in a runtime production setup.

**Personalization** and **Contextualization** are two popular directions for QR systems. A personalized system such as Cho et al., 2021 tends to incorporate diverse affinities and personal preferences to provide individually tailored user experience in a single unified system. Contextualization attempts to utilize multi-turn queries rather than only leveraging single-turn information. Some previous stud-

ies (Wang et al., 2021) have shown the benefits by leveraging the dialogue context and user-device interaction signals.

Entities have been shown to be a strong indicator of text semantics. Since queries in our dialogue system are typically short sentences, entities are even more important in this scenario. Most existing QR approaches mentioned above rephrase query utterances entirely. Although some existing works focus on specific categories like coreference resolution or entity omission (Su et al., 2019; Tseng et al., 2021), none of them has a particular emphasis on the correction of erroneous entities.

## 2.2 Entity Linking

Another related thread towards our task is entity linking. Entity linking task aims to link mentioned entities with their corresponding entities in a knowledge base. In a retrieval-based QR system which focuses on entity correction, we could adopt similar methods in entity linking area. BLINK(Wu et al., 2019) designs a two-stage retrive-rerank framework based on pre-trained deep transformers. The following work ELQ (Li et al., 2020) uses a biencoder to jointly perform mention detection and linking in one pass and also shows good improvement in latency metrics which is quite important in a production settings. Our task is more challenging than entity linking because the input utterance is noisy with incorrect entities and the lack of textual descriptions of each entity.

## 3 Problem Setup and Solution Design

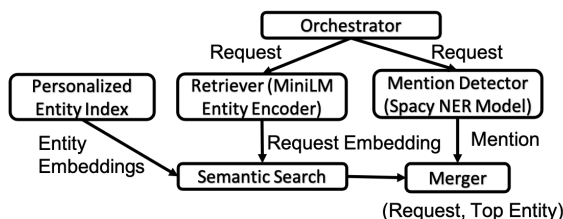The overall architecture of the PENTATRON system is described in Figure 3.



Figure 3: For a given user, the input request string from the PENTATRON orchestrator is processed by a transformer model and also by a named entity recognition model, both trained on historical user requests, to encode the request and extract mentions, respectively. A semantic search is applied on the request embeddings and the precomputed entity embeddings of the user to find the best match following which, post-processing is applied to feed the result into downstream components.

## 3.1 Entity Correction in Query Rewriting

We consider a dataset of $M$ multi-turn dialogue sessions: $\{S_t\}_{t=1}^T$. $S$ is a set of $T$ turns in chronological order: $S = \{(q_t, r_t)\}_{t=1}^T$. Here $t$ is the index of turn and each turn consists of a pair $(q_t, r_t)$, where $q_t$ denotes the user's query utterance and $r_t$ denotes the device's response utterance. The sessions are selected so that the source query $q_{T-1}$ contains one erroneous entity and $q_T$, which has the correct form of entity $e$, is the rephrase of the previous turn. More details about the data selection is described in Section 4.1. Our prediction goal is formulated as:

$$\widehat{e} = \underset{e}{\arg\max}\, \mathbf{P}(e \mid \{S_t\}_{t=1}^{T-2}, q_{T-1}) \quad (2)$$

$$\widehat{q}_T = g(q_{T-1}, \widehat{e}) \quad (3)$$

We flatten the previous dialogue turns $\{S_t\}_1^{T-2}$ and the source query $q_{T-1}$ into a single sequence to feed into the encoder, as shown in Figure 1. Since the only difference between $q_{T-1}$ and $q_T$ is whether we have the correct form of $e$, the final rewrite is generated based on source query $q_{T-1}$ and entity prediction $\hat{e}$ through a simple replacement function $g$.

## 3.2 Personalized Entity Index

We build an personalized entity index for each user to leverage individual interaction history by aggregating users' frequent entities in past 30 days[1]. The entities include song names or artists that users frequently listened to, nicknames of users' intelligent devices and so on.

This index serves as the retrieval candidate pool during inference time. The candidate embeddings are cached. We implement a two-stage in-memory index that has a map of users to their specific entities along with the embeddings corresponding to the union of entities across all users. This is done for memory efficiency reasons so that we avoid the overhead caused by the redundancy of duplicate entities across different users.

## 3.3 Modeling

We use a bi-encoder architecture based on MiniLM (Wang et al., 2020) for jointly encoding the queries and the entities (Humeau et al., 2019). The weights are shared for memory footprint savings and serving cost reduction. Note, we also try asymmetric query and candidate entity encoders;

---

[1]All user information is in a de-identified format.

however, we observe only a marginal performance improvement of less than $1\%$. We use a batch size of $128$ and train it on p3.2x-large GPU instances acquired on AWS cloud. AdamW (Loshchilov and Hutter, 2017) is our optimizer of choice.

For detecting mentions in the input query, we use a Spacy named entity recognition model trained on historical user queries containing entity strings from different domains.

### 3.4 Optimization Objectives

A combination of both hard negatives (Gillick et al., 2019) and in-batch random negatives improve the performance of large-scale natural language reasoning systems. We use the multiple negatives ranking loss (Henderson et al., 2017) for the primary task. We take a metric learning approach (Hadsell et al., 2006) to the auxiliary task, ie, we use the contrastive loss here.

**Inference:** The semantic search function which is used in primary retrieval task computes $s_i = \cos(f(q), f(e_i))$ for $i \in [k]$ where $e_i \in E_V$ are the top-$k$ entities retrieved from the personalized index (sorted by the relevance score in descending order) and $q$ refers to the query (with or without context). We configure our system to be activated on the threshold conditions, $s_1 > \tau_1$ and $s_2 < \tau_2$, to make sure the top-2 entities are sufficiently far apart to avoid any ambiguous predictions.

**Training:** The encoder model of the PENTA-TRON system is trained with the primary task of entity prediction, which we maximize the similarity score between the user query (with or without context) and the target correct entity. Consider a batch of $N$ samples. The loss of the primary task is given by:

$$\mathcal{L}_E = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s_i)}{\sum_{j=1}^{N} \exp(s_j)} \quad (4)$$

In the above formula, we only take in-batch random negatives into consideration. We will also discuss the utilization of hard negatives later in this paper.

We adopt an auxiliary task during training to have an implicit clustering effect of the query embeddings based on target domain. For this task, we want to push source queries targeting to the same domain close to each other and source queries targeting to the different domain away from each other.

For $N$ randomly selected pairs of queries (indexed by $i$ and $j$) from a batch, the loss of the auxiliary task is the contrastive loss given by:

$$\mathcal{L}_D = \frac{1}{N} \sum_{(i,j)} \mathbf{1}\{h^D(q_i) = h^D(q_j)\}.$$
$$\|f(q_i) - f(q_j)\|^2$$
$$+ \frac{1}{N} \sum_{(i,j)} \mathbf{1}\{h^D(q_i) \neq h^D(q_j)\}.$$
$$\max(0, \lambda - \|f(q_i) - f(q_j)\|^2) \quad (5)$$

The margin parameter $\lambda$ is set as $0.75$. Here, $h^D$ denotes the domain extracted from the NL hypothesis of the target (final) dialogue turn.

**Multi-task Formulation:** The final loss is computed as $\mu\mathcal{L}_E + (1 - \mu)\mathcal{L}_D$ where $\mu \in (0, 1]$. Specifically, we build different versions of PENTA-TRON by setting $\mu = 1$ and $\mu = 0.5$.

We train two single-task models which are used as the non-contextual and contextual baselines respectively. The non-contextual baseline model uses the source query as input and the rewrite entity as output. The contextual baseline model uses the full context (truncated to maximum allowable length of $256$) as input and the rewrite entity as output.

Furthermore, we train another five versions PENTATRON with multi-task settings. We also investigate the usage of task markers similar to the approach in Maillard et al., 2021. The (hard) prompts are added as special tokens [REWRITE] and [DOMAIN] before the corresponding input during training. More details are presented in Table 1

**Hard Negatives Mining:** First, we use bm25 (Robertson et al., 2009) to mine hard negatives from the candidate pool, which shows minor improvement. Hence, we adopt a two-pass method to compute hard negatives. In the first pass, we use a model trained with random negatives to perform inference on a disjoint "second" training set to obtain entity predictions. In the second pass, we continue training the previous baseline model checkpoint and take into account the wrong predictions as hard negatives.

## 4 Experiments

### 4.1 Training and Test Data

Our data is derived from the logs of a commercial voice assistant and we process the data with strict privacy standards so that users are not identifiable.

| Model | Primary Task | Auxiliary Task | Exact Match (Relative) |
|---|---|---|---|
| DPR-EC | Non-contextual | None | 0.0 [Baseline] |
| PENTATRON-N | Non-contextual | None | +432.11% |
| PENTATRON-NN | Non-contextual | Non-contextual | +438.86% |
| PENTATRON-NC | Non-contextual | Contextual | +442.76% |
| PENTATRON-NNP | Non-contextual | Non-contextual with prompt | +453.82% |
| PENTATRON-NCP | Non-contextual | Contextual with prompt | +454.47% |
| PENTATRON-C | Contextual | None | +484.14% |
| PENTATRON-CC | Contextual | Contextual | +500.97% |

Table 1: List of all model settings and their performance numbers (relative, with respect to the baseline, DPR-EC). The primary task is entity prediction using the multiple negative ranking loss with a batch size of $128$ and the auxiliary task uses the online contrastive loss with a margin of $0.75$. We apply the state-of-the-art retrieval model, DPR (Karpukhin et al., 2020), to train a dual BERT architecture, DPR-EC, for entity correction as the baseline, i.e., without utilizing personal and contextual information.

We sample multi-turn dialogue sessions between English-speaking users and devices in a time period of one month, in May-June 2022, from all over the United States. A defect detection model similar to (Gupta et al., 2021) and rule-based filters are applied to find dialogue sessions whose last two turns of user query are rephrase pairs. Rule-based filters are using edit-distance and time gap between utterance pairs similar to (Cho et al., 2021). Since our work has a particular emphasis on the correction of erroneous entities, we also utilize the NL hypothesis of the rephrase pairs to get such cases. For simplicity, we consider data with only a single erroneous entity as the target to be predicted. It is straightforward to generalize our system to the multiple entities case.

We sample the test set and keep only sessions wherein a retrieval-based system such as (Cho et al., 2021), which rephrases query utterances entirely, couldn't solve. For training, a sample of two million utterances was extracted. Also, some (completely generic) example dialogs extracted from critical data are reported in the paper (Table 2).

Figures 4 and 5 summarize the keys data statistics on the training and test sets. This gives us an insight into how transformer models stand to benefit from longer sequences in our application since they are parameterized by and compute second-order statistics.

## 4.2 Evaluation Metrics

We utilize the harshest metric to evaluate our system namely, the Exact Match (EM). This score is 1 if the predicted rewrite exactly matches the labeled rephrase, and is 0 otherwise. We use the same
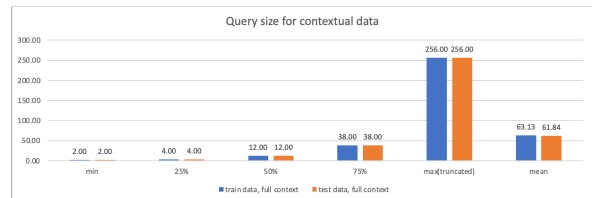


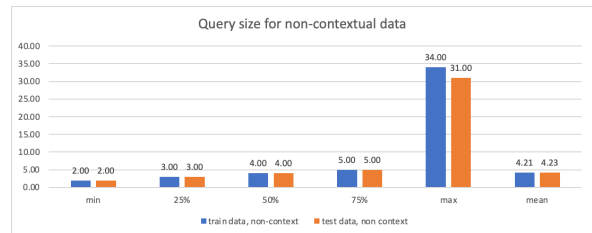Figure 4: Query length statistics of contextual training and test data.



Figure 5: Query length statistics of non-contextual training and test data.

threshold $\tau_1$ and $\tau_2$ for all the proposed PENTATRON models. The threshold is experimentally set up to keep the balance between opportunities and potential risks in real production.

## 4.3 Observations and Case-studies

Table 1 shows the main results of our different versions of systems. The experimental result is consistent with our intuition. Since the pipeline has also actually been run on live traffic, through an A/B experiment (section 4.4), the baselines were created for the purpose of this paper. All the PENTATRON models benefit from a personalization settings and outperform a global-wise retrieval model DPR-EC by a large margin. Among different settings of PENTATRON, it's obvious that both contextualiza-

| | [USER] Turn on ben's light. | [USER] Play calen playlist. |
|---|---|---|
| Dialogue Context | [DEVICE] I'm sorry I couldn't find the device. [USER] Turn on benny's light. [DEVICE] Okay. | [DEVICE] I could not find that on Amazon Music. [USER] Play scars. [DEVICE] Here's Scars , by James Bay , on Amazon Music. |
| User Query | Turn ben's light on pink | Play playlist karen |
| Rewrite Label | Turn benny's light on pink | Play playlist callen |
| DPR-EC | Turn **brecken's light** on pink ✗ | Play playlist **cameron** ✗ |
| PENTATRON-N | Turn **britney's light** on pink ✗ | Play playlist **carrie** ✗ |
| PENTATRON-C | Turn **benny's light** on pink ✔ | Play playlist **carrie** ✗ |
| PENTATRON-CC | Turn **benny's light** on pink ✔ | Play playlist **callen** ✔ |

Table 2: Two examples to showcase the importance of full contextualization and personalization.
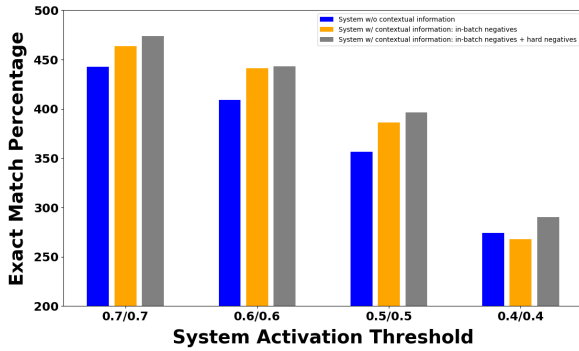


Figure 6: In this figure, we illustrate the importance of contextual information and training with hard negatives in boosting the performance of our system.
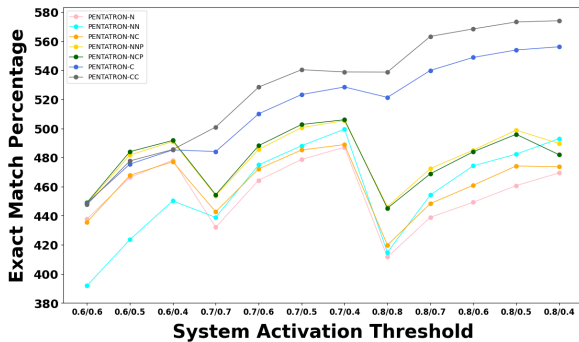


Figure 7: Performance of different versions of PEN-TATRON with respect to different system activation thresholds $\tau_1$ and $\tau_2$.



Figure 8: Demonstrating the value of contextual information with appropriate multitasking design.

are equal since the top-2 retrieved entities are semantically very similar and the model finds it difficult to disambiguate. However, with the contextual information, we see consistent improvements in accuracy as we tighten thresholds.
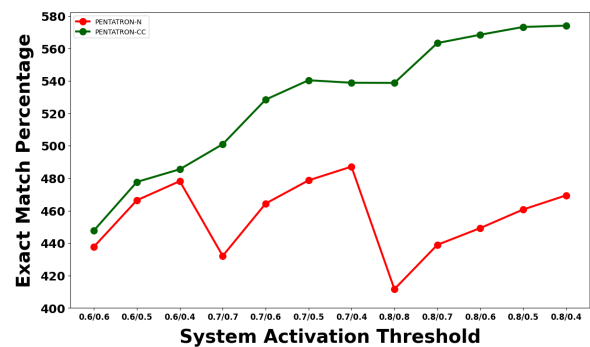
tion and multi-tasking bring further improvement. There is also some gain by adding task markers in the multi-task settings.

Figure 6 presents an ablation that shows the benefits of hard negative sampling. To further stress test our system, we also swept over different thresholds, summarized in Figure 7. We could notice that the general trend is consistent using different thresholds.

In sweeping across thresholds in our empirical studies (Figure 8), we observe interesting trends. In particular, that when $\tau_1 = \tau_2$, the personalized model that does not utilize contextual information suffers from noisy predictions when the thresholds

We illustrate the benefits of our approach on a generic dialog in Table 3. In the left example from *HomeAutomation* domain, the device name in the source query is incorrect which will make this task-oriented dialogue system fail. PENTATRON-C and PENTATRON-CC could generate the correct rewrite by leveraging dialogue context and user's personalized index which contains user's registered device name. A similar trend can be observed in the right example from Music domain. Besides, the right example also illustrates the benefits from multi-task learning by comparing the prediction from PENTATRON-C and PENTATRON-CC. Both the video name 'carrie' and playlist name 'callen' exist in user's personalized index. With the help of contrastive representation learning, PENTATRON-CC could learn to retrieve a Music domain entity which is the correct one here.

**Visualization:** We analyze the benefits of our design using t-SNE (Van der Maaten and Hinton, 2008). The results are presented in Figures 9 and 10. We clearly observe that multi-tasking enables domain disambiguation via implicitly clustering the

queries by domains, thus contributing positively to entity prediction accuracy and, in turn, improving the query rewrite quality. In particular, we observe that Music, Video and Knowledge domains immensely benefit from multi-tasking.



Figure 9: In the absence of the auxiliary task, queries across domains are interspersed which leads to lower accuracy due ambiguity in the rewrite domain. Here, the blue cluster denotes Knowledge domain queries, the orange cluster denotes Music domain queries and the green cluster denotes Video domain queries.
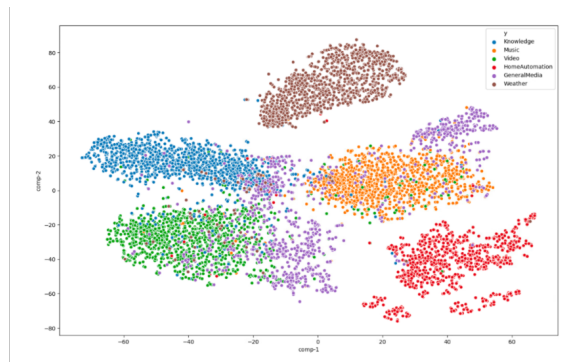


Figure 10: Multi-tasking to predict the rewrite domain, in addition to predicting the correct entity, leads to higher accuracy due to domain disambiguation arising from the implicit clustering effect.

## 4.4 Online Performance

**A/B Experimentation:** At the time of writing this, we deployed a static (request, rewrite) lookup table computed using PENTATRON-N to serve real users. With a $p$-value $< 0.05$, we observe a significant improvement, of 47.5%, in the user experience measured using the model-based (Gupta et al., 2021) assessment used for dataset selection in Section 4.1 on the treatment group as compared to the control group. Moreover, other friction metrics such as the turn error rate have improved over 40%

throughout the A/B duration. Successive version upgrade deployments are ongoing.

**Latency:** To investigate the deployment in a real-time inference service, we performed extensive load tests implemented with a Flask endpoint. We store all objects in the main memory. On a c5.9x-large instance on AWS cloud, at 120 queries per second hitting the PENTATRON system, we observed a P90 latency of less than 30ms for the end-to-end execution.

## 5 Conclusions and Future Directions

In this work, we build a system called PENTATRON which significantly improves user experience in intelligent devices by operating on entities and reducing friction in multi-turn dialogues. There are several future directions we plan to work on, including operationalizing large-scale unbiased personalized and context-aware systems, and designing self-learning (Ponnusamy et al., 2020; Roshan-Ghias et al., 2020) using techniques such as reinforcement learning. We also plan to investigate the utility of a multi-level index to improve entity coverage and mitigate the cold-start problem for new customers. Dynamic index building and deployment in low-latency applications is an ongoing direction.

## Limitations

Our system has the following limitations. Though personalization offers great benefits, the coverage of desired entities in our historical index due to personalization is typically limited. Specifically, we observe only 20% coverage in our empirical studies. This can alleviated using a multi-level index involving clusters of users. We have initial results on this approach and plan to compile that in future work.

Next, natural language based prompts should further improve our system. However, very long sequence length has concerns with respect latency and memory on CPU-deployed solutions. A potential solution to this is to consider low-rank factorization in the attention design.

Finally, in production deployments, large-scale in-memory index for multiple locales poses cost challenges. A separate study is warranted to study hybrid storage mechanisms and high performance cache design.

## Ethics Statement

To the best of our knowledge, our work is ethical and has a positive impact on society and human well-being. In particular, we take pride in emphasizing that we handle customer confidentiality and privacy with critical care. Its design principles are unbiased.

## References

Zheng Chen, Xing Fan, Yuan Ling, Lambert Mathias, and Chenlei Guo. 2020. Pre-training for query rewriting in a spoken language understanding system. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings.*

Eunah Cho, Ziyan Jiang, Jie Hao, Zheng Chen, Saurabh Gupta, Xing Fan, and Chenlei Guo. 2021. Personalized search-based query rewrite system for conversational ai. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 179–188.

Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to attend, copy, and generate for session-based query suggestion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1747–1756.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Xing Fan, Eunah Cho, Xiaojiang Huang, and Chenlei Guo. 2021. Search based self-learning query rewrite system in conversational ai. *In 2nd International Workshop on Data-Efficient Machine Learning (DeMaL).*

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. *arXiv preprint arXiv:1909.10506.*

Saurabh Gupta, Xing Fan, Derek Liu, Benjamin Yao, Yuan Ling, Kun Zhou, Tuan-Hung Pham, and Chenlei Guo. 2021. Robertaiq: An efficient framework for automatic interaction quality estimation of dialogue systems. In *2nd International Workshop on Data-Efficient Machine Learning (DeMaL).*

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652.*

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969.*

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906.*

Chanwoo Kim, Dhananjaya Gowda, Dongsoo Lee, Jiyeon Kim, Ankur Kumar, Sungsoo Kim, Abhinav Garg, and Changwoo Han. 2020. A review of on-device fully neural end-to-end automatic speech recognition algorithms. In *2020 54th Asilomar Conference on Signals, Systems, and Computers*, pages 277–283. IEEE.

Belinda Z Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. *arXiv preprint arXiv:2010.02413.*

Yiding Liu, Weixue Lu, Suqi Cheng, Daiting Shi, Shuaiqiang Wang, Zhicong Cheng, and Dawei Yin. 2021. Pre-trained language model for web-scale retrieval in baidu search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3365–3375.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101.*

Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wentau Yih, Barlas Oğuz, Veselin Stoyanov, and Gargi Ghosh. 2021. Multi-task retrieval for knowledge-intensive tasks. *arXiv preprint arXiv:2101.00117.*

Claudio Pinhanez, Paulo Cavalin, Victor Henrique Alves Ribeiro, Ana Appel, Heloisa Candello, Julio Nogima, Mauro Pichiliani, Melina Guerra, Maira de Bayser, Gabriel Malfatti, et al. 2021. Using meta-knowledge mined from identifiers to improve intent recognition in conversational systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7014–7027.

Pragaash Ponnusamy, Alireza Roshan Ghias, Chenlei Guo, and Ruhi Sarikaya. 2020. Feedback-based self-learning in large-scale conversational ai agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13180–13187.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Alireza Roshan-Ghias, Clint Solomon Mathialagan, Pragaash Ponnusamy, Lambert Mathias, and Chenlei Guo. 2020. Personalized query rewriting in conversational ai agents. *arXiv preprint arXiv:2011.04748*.

Ruhi Sarikaya. 2022. Intelligent conversational agents for ambient computing. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 5–5.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance rewriter. *arXiv preprint arXiv:1906.07004*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Bo-Hsiang Tseng, Shruti Bhargava, Jiarui Lu, Joel Ruben Antony Moniz, Dhivya Piraviperumal, Lin Li, and Hong Yu. 2021. Cread: Combined resolution of ellipses and anaphora in dialogues. *arXiv preprint arXiv:2105.09914*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Zhuoyi Wang, Saurabh Gupta, Jie Hao, Xing Fan, Dingcheng Li, Alexander Hanbo Li, and Chenlei Guo. 2021. Contextual rephrase detection for reducing friction in dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1899–1905, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.

Siyang Yuan, Saurabh Gupta, Xing Fan, Derek Liu, Yang Liu, and Chenlei Guo. 2021. Graph enhanced query rewriting for spoken language understanding system. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.