Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 385-401
DOI: https://doi.org/10.51903/jtie.v4i2.502

Federated Topic-Preference Learning for Knowledge-Grounded Chat with Differential Privacy

Meng-Ju Kuo*[1], Daren Zheng[2], Julie Hires[3]
Email: mengju.kuo0688@outlook.com
[1]*Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA*
[2]*Information Technology, Carnegie Mellon University, Pittsburgh, USA*
[3]*Computer Science, Dartmouth College, NH, USA*
*Corresponding Author

Abstract
Retrieval-augmented approaches have become central in knowledge-grounded dialogue systems, yet incorporating topical preferences remains difficult due to privacy constraints on user interaction data. This study introduces a lightweight federated topic-preference (FedTP) mechanism that models session-level preferences without centralizing raw data and uses client-level differential privacy (DP). Using the Topical-Chat dataset (8,628 conversations), each conversation is treated as a client, and evidence routing is framed as selecting relevant knowledge snippets based on dialogue context. The proposed method augments a TF-IDF relevance score with a small preference-based component derived from both local session distributions and a DP-aggregated global prior. Experimental results on 9,553 grounded test turns show a consistent but limited improvement in evidence hit rate, from 0.6167 to 0.6194. The small optimal preference weight ($\lambda = 0.005$) indicates that the preference signal mainly influences decisions when competing candidates have similar relevance scores, rather than substantially altering routing behavior. A privacy–utility analysis under Gaussian DP (ε ranging from 9.69 to 0.606, $\delta = 1e{-}5$) shows negligible changes in performance, which is expected given the large number of clients in a one-shot aggregation setting. Additional metrics remain largely stable, suggesting that the method affects selection margins rather than overall alignment. These findings suggest that federated preference aggregation can provide a modest, privacy-preserving bias for evidence routing, but its practical impact remains incremental and context-dependent.

Keywords: *Federated Learning, Differential Privacy, Knowledge-Grounded Dialogue, Topical Preference.*

I. INTRODUCTION

Retrieval-augmented generation has become a dominant paradigm for knowledge-grounded dialogue, where a model first retrieves external evidence and then generates a response conditioned on that evidence (Lewis et al., 2020; Guu et al., 2020). In open-domain settings, grounding is important not only for relevance but also for mitigating hallucinations—fluent yet unsupported generations that contradict the underlying knowledge (Maynez et al., 2020; Ji et al., 2023). At the same time, human conversations are shaped by topical preferences (Nita et al., 2025; Santi et al., 2023; Sudha & Nafeeza, 2025). Two users can be given the same pool of candidate documents yet select different facts, angles, or subtopics. Capturing such preferences can improve topic tracking and evidence routing, which may help select more appropriate citations, although it does not by itself establish factual faithfulness.

However, learning topical preferences is a privacy-sensitive problem. Preference vectors are derived from interaction logs and therefore can reveal personal interests and behavioral patterns.

Received on Juni 2025; Revised on July 2025; Accepted on August 2025; Published on August 2025

Centralizing these logs to train a single preference model raises privacy and compliance concerns, especially when the goal is personalization at the individual user or session level. Federated learning (FL) addresses this tension by training models from decentralized data without moving raw data off device (McMahan et al., 2017; Kairouz et al., 2019). Yet FL alone does not guarantee privacy: model updates may leak information about local datasets. Differential privacy (DP) provides a formal protection framework by bounding what an adversary can infer about an individual client's contribution (Dwork & Roth, 2014). Client-level DP is particularly relevant for FL because each client can correspond to a user or a session (Geyer et al., 2017).

This paper focuses on a narrow component of knowledge-grounded chat: evidence routing. Rather than training a full end-to-end generator, we study whether session-level preference vectors can nudge the selection of knowledge snippets to cite for each response. We use the Topical-Chat dataset, which was constructed to elicit knowledge-grounded open-domain conversations using curated reading sets (Gopalakrishnan et al., 2019). In Topical-Chat, each grounded turn is annotated with the knowledge source the speaker used (e.g., one of three "fun fact"/Wikipedia sources, FS1–FS3), enabling direct evaluation of routing quality.

We propose Federated Topic-Preference (FedTP) learning: each conversation is treated as a client that locally estimates a preference distribution over evidence topics from observed grounded turns. A server aggregates these client preferences into a global prior μ using clipping and Gaussian noise to satisfy client-level DP. At inference time, the router combines a standard TF-IDF relevance score with a small preference-based score derived from μ and the client's local preference. This design is deliberately lightweight: it isolates the question of whether DP federated aggregation can provide a small but useful inductive bias for routing, and it can be integrated with any downstream generator that cites retrieved evidence. Because Topical-Chat conversations are independent and do not correspond to persistent users, the resulting setup should be interpreted as session-level aggregation rather than long-term user personalization.

Our study makes three concrete contributions: (1) We implement a reproducible FedTP pipeline on Topical-Chat and report detailed dataset statistics, hyperparameters, and metrics. (2) We empirically compare a TF-IDF-only router against global, local, and mixed (federated) preference augmentation, showing a small but consistent gain in evidence hit rate on the full test set. Because the best preference weight is very small, we interpret this gain as a tie-breaking effect rather than a large change in routing behavior. (3) We quantify the privacy–utility tradeoff by varying the DP noise multiplier and computing the corresponding privacy budget ε, and we measure its effect not only on routing accuracy but also on an intrinsic lexical-support indicator based on evidence–response token overlap.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 385-401
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

II. LITERATURE REVIEW

A. Knowledge-Grounded Dialogue and Evidence Citation

Early open-domain dialogue systems relied heavily on sequence-to-sequence generation trained on conversational corpora, which often produced generic or inconsistent responses (Vinyals & Le, 2015). To improve informativeness and factuality, several benchmarks introduced explicit knowledge grounding. Wizard of Wikipedia pairs dialogue with Wikipedia passages and encourages models to select and use supporting sentences (Dinan et al., 2019). Topical-Chat similarly provides curated reading sets and annotates which knowledge source each speaker used, emphasizing natural and engaging conversation rather than strict question answering (Gopalakrishnan et al., 2019). More recent work popularized retrieval-augmented generation, where a retriever selects documents and a generator conditions on them (Lewis et al., 2020; Guu et al., 2020). Retrieval-based knowledge integration has also been shown to reduce hallucination in knowledge-grounded conversation settings (Shuster et al., 2021). Although our experiments focus on routing rather than generation, these lines of work motivate treating evidence selection as a first-class component of grounded chat.

B. Personalization and Topical Preferences

Personalization has been studied in dialogue through persona conditioning and user embeddings. Persona-based dialogue introduces explicit persona sentences that models must incorporate, demonstrating that conditioning can improve consistency (Zhang et al., 2018). In open-domain assistants, user preferences are often implicit (e.g., recurring topics) rather than explicit statements of persona. Preference learning is also central in recommendation systems, where topic and taste vectors are learned from user histories. In the context of grounding, preferences can affect which documents are retrieved or which facts are selected from the same set of documents. In this paper, we model such variation at the session level rather than as long-term user personalization.

C. Federated Learning for Personalization

Federated learning was introduced to enable training across large numbers of clients while keeping raw data local, with FedAvg as a canonical aggregation algorithm (McMahan et al., 2017). Because client data are typically non-IID, personalization is a recurring theme in FL (Kairouz et al., 2019). Several approaches split a model into global and local components, such as learning shared representations while keeping a personalization head on device (Arivazhagan et al., 2019). Other approaches regularize local updates toward the global model to improve robustness to heterogeneity (Li et al., 2020), and more recent work studies explicit personalized memorization under heterogeneous client distributions (Marfoq et al., 2022). While most FL personalization work targets classification or next-word prediction (Hard et al., 2018), the same

principles apply to routing policies in dialogue: clients can learn local routing biases while benefiting from a global prior.

D. Differential Privacy in Federated Settings

Differential privacy provides a formal guarantee that bounds the change in an algorithm's output when a single client's data are added or removed (Dwork & Roth, 2014). In deep learning, DP is commonly implemented by clipping gradients and adding noise (Abadi et al., 2016). For FL, client-level DP can be achieved by clipping each client update and adding noise to the aggregated update, which protects the presence or absence of an entire client (Geyer et al., 2017). Rényi DP provides tighter accounting for repeated applications of the Gaussian mechanism (Mironov, 2017), and multiple works have explored DP for language modeling and next-word prediction (McMahan et al., 2018). Our approach follows the same core mechanism, clipping and Gaussian noise, but applies it to preference vectors rather than model gradients.

E. Hallucination and Grounding Evaluation

Evaluating hallucination is challenging because it depends on external knowledge and context, and factual consistency cannot be fully recovered from surface overlap alone (Maynez et al., 2020; Ji et al., 2023). In dialogue, grounding can be evaluated intrinsically by checking whether the selected evidence matches annotations (Dinan et al., 2019; Gopalakrishnan et al., 2019) and extrinsically by measuring consistency between responses and evidence. Automatic metrics such as BLEU and ROUGE were originally introduced for translation and summarization (Papineni et al., 2002; Lin, 2004) and have been used in dialogue despite their known limitations (Liu et al., 2016). Because this paper isolates routing, we report a simple support rasio, the fraction of response tokens that also appear in the selected evidence snippet, as an intrinsic lexical-support indicator. We do not treat this ratio as a direct factuality or hallucination metric, since high overlap can still be unfaithful and low overlap does not necessarily imply hallucination (Maynez et al., 2020; Ladhak et al., 2022).

III. RESEARCH METHOD

This section describes the dataset, the federated preference-learning procedure, the DP aggregation mechanism, and the evaluation protocol. Figure 1 summarizes the overall pipeline.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 385-401
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

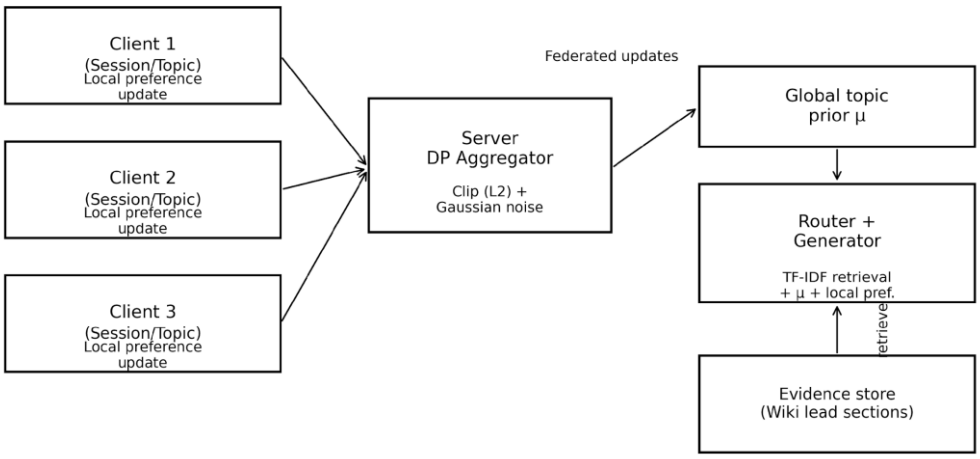

Figure 1. FedTP Pipeline: Client-Side Preference Estimation, DP Aggregation, and Preference-Augmented Routing

A. Dataset

We use Topical-Chat, a knowledge-grounded open-domain conversation dataset built from curated reading sets and human–human dialogues (Gopalakrishnan et al., 2019). We use the Hugging Face JSONL splits and treat each conversation as a federated client. Each dialogue turn includes an annotation "knowledge_source" indicating which source the speaker used (e.g., FS1–FS3, AS1–AS4, or Personal Knowledge). We evaluate only FS-labeled turns because they correspond to explicit evidence candidates in the reading set. Table 1 reports split statistics. The training set contains 8,628 conversations (188,378 turns). Across splits, approximately 78% of turns are FS-grounded, which makes routing evaluation well-defined.

Table 1. Topical-Chat Split Statistics (Conversations and Turns)

Split	Conversations	Turns	Avg. turns/conv	Grounded turns (FS)	Grounded %	Avg. grounded turns/conv
Train	8628	188378	21.83	146514	77.78	16.98
Valid-Freq	539	11681	21.67	8904	76.23	16.52
Valid-Rare	539	11692	21.69	9131	78.1	16.94
Test-Freq	539	11760	21.82	8551	72.71	15.86
Test-Rare	539	11770	21.84	8995	76.42	16.69

B. Knowledge-Source Distribution

Table 2 summarizes the distribution of the annotated knowledge sources in the training split. FS1–FS3 dominate the dataset, while article-sentence sources (AS1–AS4) and Personal Knowledge appear less frequently. Because our router operates over the FS candidate set, we filter evaluation turns to those with at least one FS label.

Table 2. Training Split knowledge_source Label Distribution

Knowledge source	Count	Percent
FS1	59907	26.92
FS2	50220	22.56
FS3	41270	18.54
Personal Knowledge	50529	22.7
AS1	10485	4.71
AS2	4281	1.92
AS4	2638	1.19
AS3	3229	1.45

C. Client Statistics

Conversations are short but non-trivial: most contain 20–23 turns, with a mean of 21.83 turns. Grounded turns per conversation average 16.98 in training and 16.28 in test, but some conversations contain no FS-labeled turns. Table 3 reports distributional statistics for conversation length and grounded-turn counts.

Table 3. Conversation Length and Grounded-Turn Statistics (Percentiles)

Split	Metric	mean	std	p10	p25	p50	p75	p90	min	max
Train	Conversation turns	21.83	1.75	21.0	21.0	21.0	22.0	23.0	20.0	53.0
Train	Grounded turns (FS)	16.98	4.59	10.0	14.0	18.0	20.0	21.0	0.0	33.0
Test (Freq+Rare)	Conversation turns	21.83	1.93	21.0	21.0	21.0	22.0	23.0	20.0	51.0
Test (Freq+Rare)	Grounded turns (FS)	16.28	4.96	9.0	13.0	18.0	20.0	21.0	0.0	31.0

D. Evidence Store

Each conversation is associated with a small reading set for each speaker, including three Wikipedia-derived lead sections (FS1–FS3). The Hugging Face split files reference these lead sections using unique identifiers for shortened or summarized variants. We reconstruct the evidence text using the official mapping provided in the Topical-Chat repository (Gopalakrishnan et al., 2019). The resulting evidence store contains 516 unique wiki lead sections (261 shortened and 255 summarized). Table 4 summarizes evidence statistics.

Table 4. Evidence-Store Statistics (Unique Wiki Lead Sections)

Evidence type	Count	Avg. tokens
Shortened wiki lead	261	135.05
Summarized wiki lead	255	146.78
Total (union)	516	140.85

E. Client Definition and Local Preference

We treat each conversation as a federated client. For a given client c, we define a preference vector $\mathrm{p}_c$over the global evidence-topic space, which consists of 516 Wikipedia lead sections. The vector $\mathrm{p}_c$is estimated from the first half of the conversation as a normalized histogram of the

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 385-401
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

evidence topics referenced by FS-labeled turns. Formally, for topic i, the preference value is defined as

$$p_{c,i} = \frac{\text{count}_c(i)}{\sum_j \text{count}_c(j)},$$

where $\text{count}_c(i)$denotes the number of times topic $i$appears in the FS-labeled turns of client c. If a conversation contains no FS-labeled turns in its first half, $\mathbf{p}_c$is set to a uniform distribution over all topics. Because conversations in Topical-Chat are independent and do not correspond to persistent users, this formulation should be interpreted as session-level preference estimation rather than user-level personalization

F. Federated Aggregation of a Global Prior

We aggregate client preferences into a global prior μusing a FedAvg-style mean:

$$\mu = \frac{1}{N}\sum_{c=1}^{N} \mathrm{p}_c,$$

where $N$is the number of training conversations. This aggregation yields a dataset-level topic prior that captures which evidence topics are common across clients. In the present study, the aggregation is performed once on the training set and released once; accordingly, the reported privacy budget applies to a one-shot aggregation setting.

G. Client-Level Differential Privacy

To protect client contributions, we apply client-level DP during aggregation. Each $\mathbf{p}_c$is L2-clipped to norm $C(C = 1.0)$ before aggregation, and we add isotropic Gaussian noise to the summed vector:

$$\mathrm{S} = \sum_{c=1}^{N} \text{clip}\,(\mathrm{p}_c, C) + \mathcal{N}(0, \sigma^2 C^2 \mathrm{I}).$$

The released prior is

$$\hat{\mu} = \text{normalize}\left(\max\left(\frac{\mathrm{S}}{N}, 0\right)\right),$$

where “normalize” projects to the probability simplex by renormalizing the non-negative components. For the released mean, the L2 sensitivity scales as C/N, so after dividing by $N$the effective noise standard deviation per coordinate is $\sigma C/N$. With $N = 8{,}628$clients, this scaling explains why the utility curve is nearly flat over the tested σvalues. For $\delta = 10^{-5}$, the corresponding privacy budget is

$$\varepsilon = \frac{\sqrt{2\ln\left(\frac{1.25}{\delta}\right)}}{\sigma}$$

for this one-shot Gaussian mechanism (Dwork & Roth, 2014; Abadi et al., 2016). Repeated aggregation rounds or client subsampling would require different privacy accounting.

H. Preference Mixing

At inference time we mix the global prior $\hat{\mu}$with the local client preference $\mathrm{p}_c$to obtain

$$\hat{\mathrm{p}}_c = \alpha\hat{\mu} + (1-\alpha)\mathrm{p}_c.$$

We use $\alpha = 0.7$(global weight), which yields a smoothed local preference vector

I. Evidence Router

Each grounded turn has three candidate evidence snippets corresponding to FS1–FS3 for the speaker. We represent evidence snippets and queries using TF-IDF with unigrams and bigrams (Salton & Buckley, 1988). The query $q_t$for turn $t$is the concatenation of the previous $k = 3$utterances. For a candidate evidence snippet e, the router computes a base relevance score

$$r(e) = \cos\ (\text{TF-IDF}(q_t), \text{TF-IDF}(e)).$$

FedTP augments this with a preference term:

$$\text{score}(e) = r(e) + \lambda \cdot \hat{\mathrm{p}}_c[e].$$

We tune λon validation and use $\lambda = 0.005$for all test experiments. Because the selected $\lambda$is very small and larger λvalues degrade validation performance, the preference component should be interpreted as a mild bias that mainly affects borderline cases rather than a dominant routing signal. The key hyperparameters and experimental settings used in this component are summarized in Table 5.

Table 5. Key Hyperparameters and Experimental Settings

Component	Setting
Evidence corpus	516 wiki lead sections (shortened+summarized)
Vectorizer	TF-IDF, ngram_range=(1,2), stop_words=english
Router base score	Cosine(query, evidence) in TF-IDF space
Context window k	3 previous utterances concatenated
Preference mixing	p_hat = 0.7*mu + 0.3*p_local
Preference weight λ	0.005 (tuned on validation)
DP clipping norm C	1.0 (L2 clip on client preference vectors)
DP mechanism	Gaussian noise on aggregated sum
DP delta δ	1e-5
Noise multipliers σ	{0.5,1,2,4,8} (plus σ=0 no DP)
Random seed(s)	0 for main results

J. Evaluation Protocol

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 385-401
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

We compute $\hat{\mu}$ from the training split and tune λon the validation splits (valid_freq + valid_rare). For evaluation, we use the test splits (test_freq + test_rare). For each conversation, we use the first half of turns to compute p_c and evaluate only FS-labeled turns in the second half, resulting in 9,553 evaluated turns.

K. Metrics

We report: (1) Evidence hit rate (EHR), equivalent to topic hit rate in this setting, defined as the fraction of evaluated turns where the router selects the same FS evidence topic as the dataset annotation. (2) Evidence–response relevance (ERR), the cosine similarity between TF-IDF(response) and TF-IDF(selected evidence), which measures intrinsic alignment between what was said and what was cited. (3) Support ratio, the fraction of response tokens that appear in the selected evidence snippet (after lowercasing and tokenization). We interpret the support ratio only as an intrinsic lexical-support indicator. It is not a direct hallucination or factuality evaluation metric, because lexical overlap does not guarantee faithfulness and low overlap does not by itself imply hallucination.

All non-DP components are deterministic given the preprocessing and split definitions. For the DP setting, the only stochasticity is the Gaussian noise used in the one-shot aggregation. We report the main tables with seed 0 and additionally repeat the DP runs over five random seeds. Across $\sigma \in \{0.5, 1, 2, 4, 8\}$, the method ranking remains stable, with the mean overall EHR staying in the narrow range of 0.61976–0.61995 and the standard deviation between 0.00005 and 0.00018.

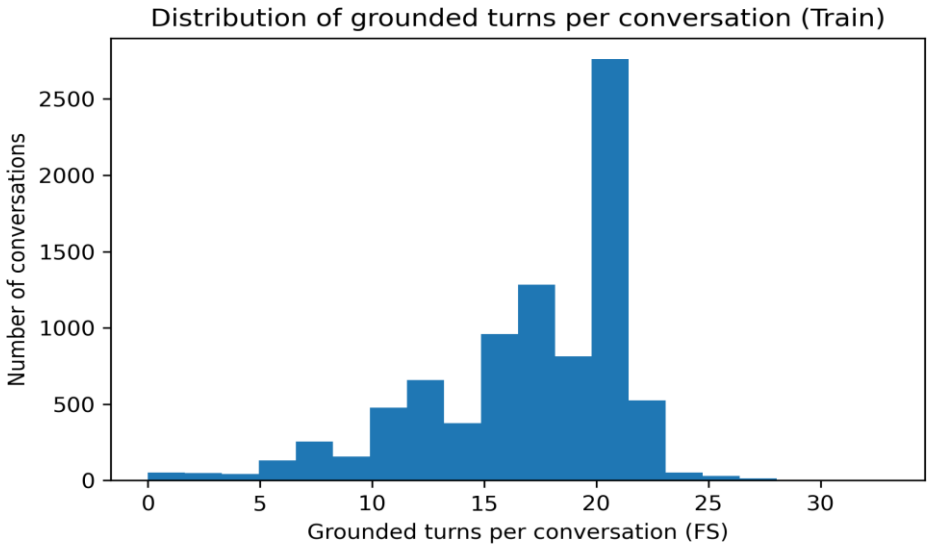

Figure 2. Histogram of Grounded Turns per Conversation in the Training Split

IV. RESULT

Figure 2 shows the distribution of grounded turns per conversation in the training split. Most conversations contain 10–22 grounded turns, which provide enough signal to estimate a local preference vector without requiring long interaction histories.

A. Hyperparameter Tuning

We tuned the preference weight λ on the validation splits by sweeping $\lambda \in \{0, 0.005, 0.01, 0.02, 0.05\}$. Table 6 reports the validation results. $\lambda = 0.005$ achieved the best evidence hit rate and was used for all subsequent experiments. The narrow optimum and the rapid degradation at larger λ show that lexical relevance remains the dominant routing signal; the preference term is useful primarily as a conservative tie-breaker.

Table 6. Validation Sweep for Preference Weight λ (α = 0.7)

λ	EHR	ERR	Support ratio
0.0	0.5816	0.0523	0.2344
0.005	0.5836	0.0524	0.2343
0.01	0.5821	0.0523	0.2343
0.02	0.579	0.052	0.2341
0.05	0.5665	0.0514	0.2333

B. Main Experimental Comparison

Table 7 compares four routers on the combined test set: (i) a TF-IDF-only baseline, (ii) a global-prior-augmented router, (iii) a local-preference-augmented router, and (iv) FedMix, which mixes global and local preferences. FedMix achieves the highest overall evidence hit rate (0.6194), compared with 0.6167 for the baseline. This +0.27 percentage-point gain is consistent but modest, so the result should be interpreted as an incremental routing improvement rather than a large practical change. ERR and support ratio remain nearly unchanged, indicating that the preference term mainly nudges which of the three candidate snippets is selected when lexical evidence is already competitive, rather than substantially altering lexical alignment between responses and snippets. Whether this gain justifies the added complexity of the federated pipeline depends on whether small improvements in borderline routing cases matter in the target deployment.

Table 7. Test results: Comparison of Routers (Overall and per Split)

Method	EHR (freq)	EHR (rare)	EHR (overall)	ERR (overall)	Support ratio (overall)
TF-IDF Router (no pref)	0.5858	0.6475	0.6167	0.0499	0.2425
Global prior + TF-IDF	0.5848	0.6517	0.6182	0.0498	0.2421
Local pref + TF-IDF	0.5936	0.6444	0.619	0.0498	0.2421
FedMix (α=0.7) + TF-IDF	0.5933	0.6454	0.6194	0.0497	0.2423

To quantify uncertainty around the observed gain, we additionally computed paired turn-level bootstrap confidence intervals over the combined test set. The 95% bootstrap confidence interval

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 385-401
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

for the absolute EHR difference (FedMix − TF-IDF baseline) is [0.0003, 0.0063]. The gain is therefore statistically distinguishable at the turn level, while remaining practically modest.

C. *Privacy–Utility Tradeoff*

Table 8 and Figure 3 show the effect of client-level DP on evidence hit rate as the privacy budget ε varies. For δ = 1e-5, σ ∈ {0.5, 1, 2, 4, 8} corresponds to ε values from 9.69 down to 0.606. The practically stronger-privacy region in this sweep is the lower-ε range (approximately 0.6–2.4), whereas σ = 0.5 (ε = 9.69) should be interpreted as weak privacy, and σ = 0 as a no-DP reference point rather than a privacy-preserving setting. In this dataset, the performance curve is nearly flat. This is not surprising: the released statistic is a clipped mean over N = 8,628 clients, so its sensitivity is C/N and the injected noise after averaging has scale σC/N. With a large N, the perturbation is small relative to the aggregated signal, so negligible utility degradation is expected. The reported ε values also apply only to this one-shot aggregation; repeated rounds or client subsampling would require tighter composition-based accounting.

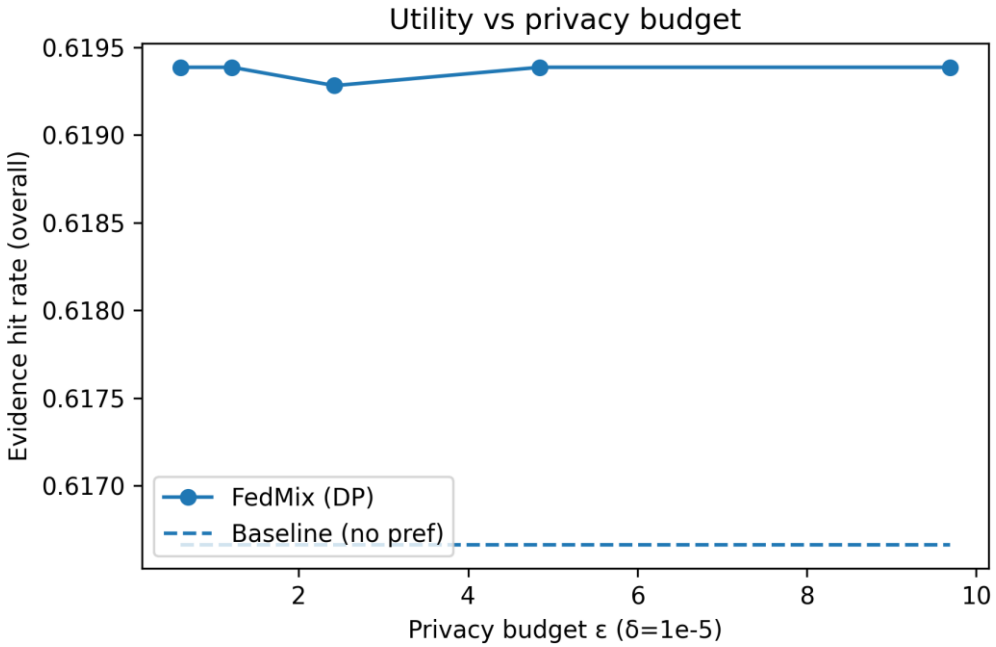

Figure 3. Evidence Hit Rate (Overall) vs. Privacy Budget ε

Table 8. DP Tradeoff on Test (FedMix): Utility vs Privacy Budget for One-Shot Aggregation

sigma	epsilon	overall_ehr	overall_gain	overall_err	overall_sr
0.0	∞	0.6194	0.0027	0.0497	0.2423
0.5	9.69	0.6194	0.0027	0.0497	0.2423
1.0	4.845	0.6194	0.0027	0.0497	0.2423
2.0	2.422	0.6193	0.0026	0.0497	0.2423
4.0	1.211	0.6194	0.0027	0.0497	0.2423
8.0	0.606	0.6194	0.0027	0.0497	0.2424

Because DP introduces Gaussian noise, we also repeated the DP runs over five random seeds. Across $\sigma \in \{0.5, 1, 2, 4, 8\}$, the method ranking remained stable, with mean overall EHR ranging from 0.61976 ± 0.00018 ($\sigma = 8$, $\varepsilon \approx 0.606$) to 0.61995 ± 0.00006 ($\sigma = 0.5$ or 1.0). This confirms that the near-flat privacy–utility curve is not an artifact of a single seed.

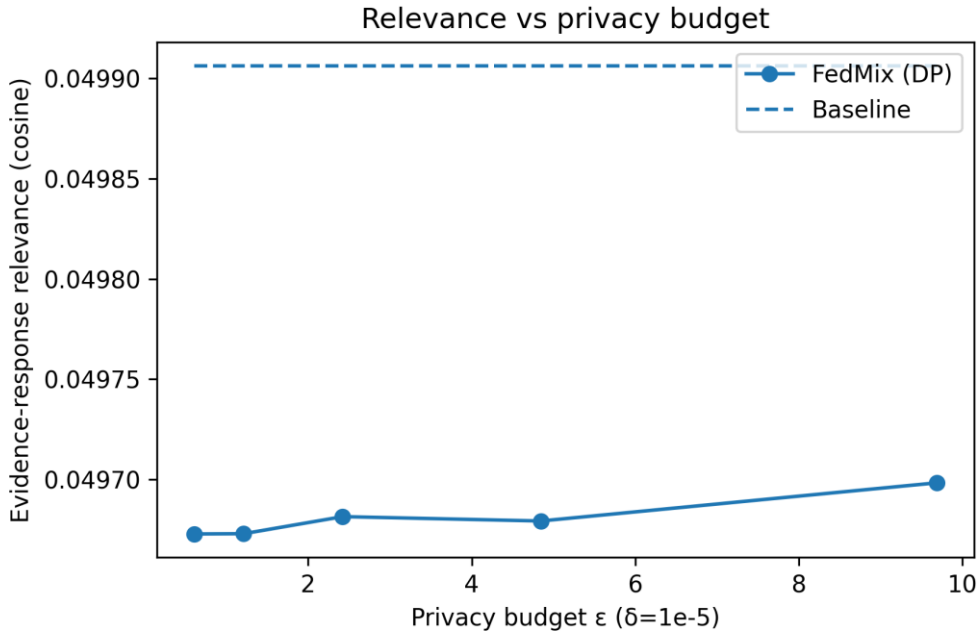

Figure 4. Evidence–Response Relevance (Cosine) vs. Privacy Budget ε

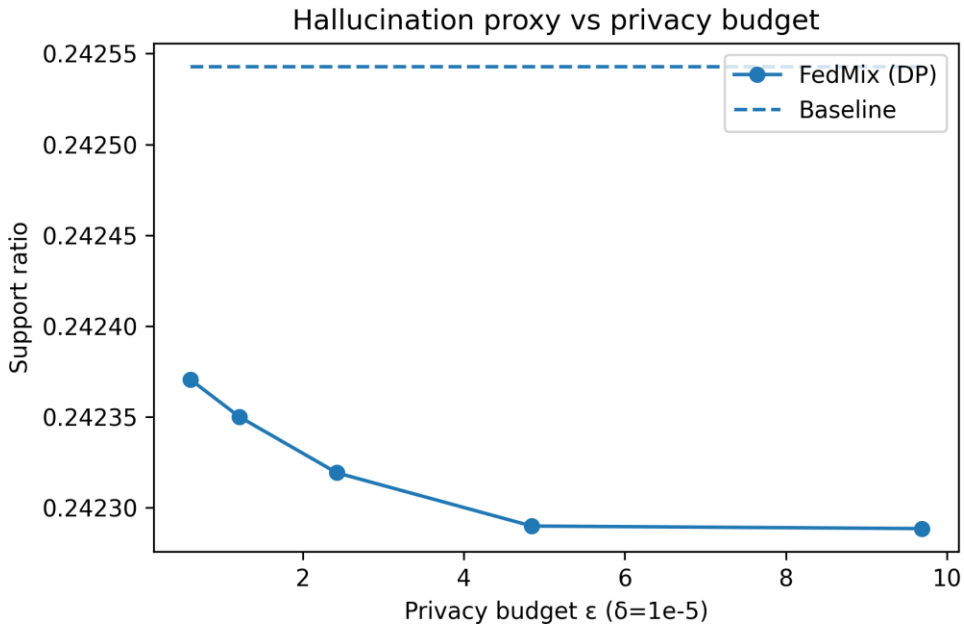

Figure 5. Support Ratio vs. Privacy Budget ε (Higher Indicates Greater Lexical Overlap with Evidence, Not a Direct Factuality Score)

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 385-401
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Figures 4 and 5 present the same DP sweep for evidence–response relevance and support ratio. FedMix slightly decreases ERR and support ratio relative to the baseline, and DP has no practical impact on either metric in this setting. As noted earlier, support ratio should be interpreted only as a lexical-support indicator, not as a direct factuality metric.

D. Frequent vs. Rare Evaluation

Table 9 reports results on the frequency-based test partitions. FedMix improves evidence hit rate on the frequent split (0.5933 vs. 0.5858) and is slightly below the baseline on the rare split (0.6454 vs. 0.6475). This pattern is expected because the aggregated prior μ is shaped primarily by commonly observed topics. As a result, the method favors routing decisions aligned with frequent topics and can smooth away some niche-topic variation. In this sense, FedMix trades a small gain on frequent-topic routing for little or no benefit on rare-topic routing.

Table 9. Test Split Results (Frequent vs. Rare) for Baseline and FedMix

Split	Method	EHR	ERR	Support ratio
Test-Freq	Baseline (TF-IDF)	0.5858	0.0403	0.2287
Test-Freq	FedMix (α=0.7, λ=0.005)	0.5933	0.04	0.2284
Test-Rare	Baseline (TF-IDF)	0.6475	0.0595	0.2564
Test-Rare	FedMix (α=0.7, λ=0.005)	0.6454	0.0594	0.2562

E. Personalization Gain Distribution

Although the average gain is small, gains are not uniform across conversations. Table 10 summarizes per-conversation changes in evidence hit rate, and Figure 6 visualizes the distribution. Most conversations are unchanged because TF-IDF relevance already resolves many turns; specifically, 948 of 1,058 conversations with evaluation turns (89.6%) are unchanged, while 60 improve and 50 worsen. This distribution is consistent with the very small λ selected on validation: the preference term does not broadly reshape routing behavior, but occasionally changes decisions in a minority of conversations where the lexical scores are likely close. A more targeted ambiguity analysis, for example, restricting attention to turns where the top-2 cosine scores are separated by only a small margin, would help quantify this tie-breaking effect more directly.

Table 10. Per-Conversation Gain Summary on the Combined Test Set

Statistic	**Value**
Conversations with eval turns	1058.0
Improved conversations	60.0
Worsened conversations	50.0
Unchanged conversations	948.0
Mean gain	0.0016
Median gain	0.0
Max gain	0.6364
Min gain	-1.0

Across the full combined test set, the preference term changed 298 of 9,553 routing decisions (3.12%). To examine where these changes occur, we isolated ambiguous turns for which the top-2 TF-IDF cosine scores differ by at most 0.005. On this subset (1,076 turns; 11.3%), the preference term flipped the selected evidence on 27.7% of turns and improved EHR from 0.3532 to 0.3820. Outside this ambiguous subset, routing decisions were unchanged in this rerun. This supports the interpretation of FedMix as a tie-breaking mechanism rather than a dominant routing signal.

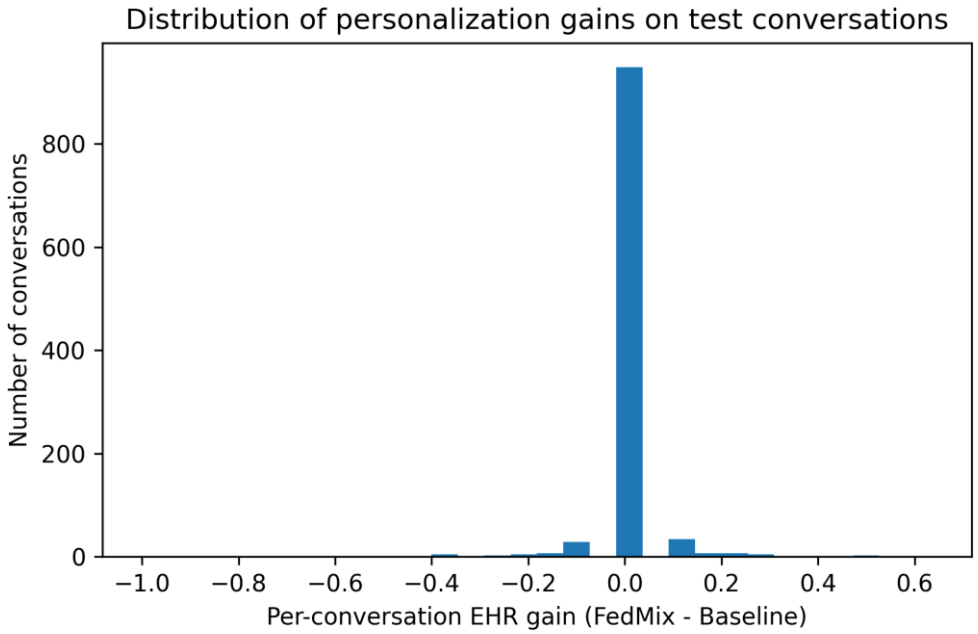

Figure 6. Distribution of per-Conversation Evidence Hit-Rate Gains (FedMix − Baseline)

V. CONCLUSION AND RECOMMENDATION

This paper evaluated a lightweight federated topic-preference mechanism for knowledge-grounded chat under client-level differential privacy. Using Topical-Chat, we treated each conversation as a client, computed local preference distributions over evidence topics, and aggregated a global prior with clipping and Gaussian noise. A simple TF-IDF router augmented with a very small preference term ($\lambda = 0.005$) improved evidence hit rate from 0.6167 to 0.6194 on 9,553 grounded test turns. We interpret this as a modest tie-breaking gain, not as a large routing improvement. Because conversations in Topical-Chat are independent and do not correspond to persistent users, the present results demonstrate session-level preference aggregation under DP rather than long-term user personalization. The DP sweep (ε from 9.69 to 0.606 at δ = 1e-5) showed negligible utility loss, which is expected in a one-shot aggregation over 8,628 clients because the mean sensitivity and post-averaging noise scale are both proportional to 1/N. Across all settings, evidence–response relevance and lexical support remained stable, suggesting that

preference learning mainly affects borderline candidate selection rather than overall response–evidence alignment.

Based on these findings, we recommend a conservative interpretation of FedTP. The method is most appropriate when a lightweight federated prior is desired and small improvements in borderline routing decisions are worth the added pipeline complexity. In its current form, the approach should not be presented as strong personalization, since the client unit is a session rather than a persistent user, and the reported privacy budget applies only to a single aggregation release. Future work should extend FedTP beyond TF-IDF to semantic retrievers or citation-aware generators while maintaining privacy accounting for repeated rounds and client subsampling.

REFERENCES

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. https://doi.org/10.1145/2976749.2978318

Arivazhagan, M., Aggarwal, V., Singh, A. K., & Choudhary, S. (2019). Federated Learning with Personalization Layers. arXiv Preprint, *arXiv:1912.00818*. https://arxiv.org/abs/1912.00818

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., & Weston, J. (2019). Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations (ICLR)*, *2019*. https://arxiv.org/abs/1811.01241

Dwork, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3), 211–407. https://doi.org/10.1561/0400000042

Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially Private Federated Learning: A Client Level Perspective. arXiv Preprint, *arxiv:1712.07557*. https://arxiv.org/abs/1712.07557

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., & Hakkani-Tür, D. (2019). Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proceedings of Interspeech 2019*, 1891–1895. https://doi.org/10.21437/Interspeech.2019-3133

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/2002.08909

Hard, A., Rao, K., Mathews, R., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., & Ramage, D. (2018). Federated Learning for Mobile Keyboard Prediction. arXiv Preprint, *arxiv:1811.03604*. https://arxiv.org/abs/1811.03604

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3591421

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2019). Advances and Open Problems in Federated Learning. arXiv Preprint, *arxiv:1912.04977*. https://arxiv.org/abs/1912.04977

Ladhak, F., Durmus, E., He, H., Cardie, C., & McKeown, K. (2022). Faithful or Extractive? On Mitigating the Faithfulness-Abstractiveness Trade-Off in Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1410–1421. https://doi.org/10.18653/v1/2022.acl-long.102

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. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2005.11401

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated Optimization in Heterogeneous Networks. In *Proceedings of Machine Learning and Systems (MLSys)*. https://arxiv.org/abs/2003.00295

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. https://doi.org/10.3115/1220355.1220364

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). How NOT to Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2122–2132. https://doi.org/10.18653/v1/d16-1230

Nita, E., Tute, K. J., & Sala, E. E. (2025). Web-Based Information System for Student Internship Data Processing Using the Agile Method. Jurnal Ilmiah Sistem Informasi, *4*(3), 571–589. https://doi.org/10.51903/7hmb1006

Marfoq, O., Neglia, G., Vidal, R., & Kameni, L. (2022). Personalized Federated Learning Through Local Memorization. In *Proceedings of the 39th International Conference on Machine Learning*, 15070–15092. https://proceedings.mlr.press/v162/marfoq22a.html

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-Efficient Learning of Deep Networks From Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html

McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2018). Learning Differentially Private Recurrent Language Models. In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1710.06963

Mironov, I. (2017). Rényi Differential Privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 263–275. https://doi.org/10.1109/csf.2017.11

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318. https://doi.org/10.3115/1073083.1073135

Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Santi, D., Harahap, R. D., & Siregar, S. U. (2023). Scientific Attitude Analysis on Students in Class XI-IPA at SMA Negeri 2 Bilah Hulu Regarding Human Blood Circulation System Material. Journal of Management and Informatics, *2*(1), 01–07. https://jmi.stekom.ac.id/index.php/jmi/article/view/13

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803. https://doi.org/10.18653/v1/2021.findings-emnlp.321

Sudha, S., & Nafeeza, S. (2025). Predicting and Inspecting Food Contamination Using AI-Based Hyperspectral Imaging. Journal of Technology Informatics and Engineering, *4*(2), 204–213. https://doi.org/10.51903/jtie.v4i2.266

Vinyals, O., & Le, Q. V. (2015). A Neural Conversational Model. *arXiv preprint arXiv:1506.05869*. https://arxiv.org/abs/1506.05869

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2204–2213. https://doi.org/10.18653/v1/p18-1196