

Probing Benchmark-Specific Regularities in Knowledge Gating via Lightweight Feature Analysis

Zilong Cheng¹, Yubo Chen^{1,2}, Jingbo Sun^{1,2},
Ge Gao^{1,2}, Bo Chen^{1,2}, Xiaobing Zhao^{1,2}

¹School of Information Engineering, Minzu University of China, Beijing 100081, China

²National Language Resource Monitoring and Research Center of Minority Languages,
Minzu University of China, Beijing 100081, China

25012081@muc.edu.cn, yubo.chen@muc.edu.cn, sunjingbo@muc.edu.cn
ggao.bj@qq.com, chenbomuc@muc.edu.cn, nmzxb_cn@163.com

Abstract

We study whether knowledge-gating benchmarks support transferable retrieval decisions or expose benchmark-specific structural regularities that can be exploited without full semantic interpretation. Knowledge gating is the binary decision of whether a dialogue system should retrieve external knowledge at a given turn. We audit this question with a lightweight logistic-regression probe over ten dialogue-metadata features: position, length, and question-form cues. Despite using only metadata and surface signals, the probe reveals non-trivial in-domain signal and markedly different structural patterns across benchmark families. Cross-benchmark transfer provides the clearest evidence: transfer within the related DSTC9/DSTC11 family is near-lossless (Macro F1 0.81–0.84; ROC-AUC $\approx$ 0.90), whereas transfer across the KETOD–DSTC benchmark-family boundary has weak ranking quality (ROC-AUC 0.41–0.48) and poor detection of knowledge-needed turns (positive-class $F1 \leq 0.22$ in one direction). Sentence embeddings and fine-tuned BERT show the same qualitative weakness in the DSTC→KETOD setting: MiniLM ranking is sub-chance (ROC-AUC 0.40–0.44), while BERT remains near chance (0.50–0.53) despite having the strongest KETOD in-domain positive F1. These results suggest that strong in-domain gating scores can partly reflect benchmark-specific regularities and should not, by themselves, be interpreted as evidence of transferable retrieval reasoning. Code and reproducibility materials are available at <https://github.com/zilongcheng582-cyber/RAGate-analysis>.

1 Introduction

Retrieval-augmented dialogue systems do not need external knowledge at every turn. A user asking whether a restaurant has outdoor seating may require retrieval; a user saying “thanks” usually does not. We use *knowledge gating* for the binary decision of whether to retrieve external knowledge at a

dialogue turn, closely related to knowledge-seeking turn detection and adaptive retrieval. This decision is central to retrieval-augmented generation (Lewis et al., 2020; Shi et al., 2023): inaccurate gating either adds unnecessary latency or omits needed evidence. Models such as RAGate (Wang et al., 2025) frame this decision as binary classification and report strong in-domain benchmark performance.

Strong in-domain performance, however, need not imply transferable gating ability. NLI models can exploit annotation artifacts—lexical overlap, negation, and hypothesis-only cues—rather than the target task (Gururangan et al., 2018; McCoy et al., 2019; Poliak et al., 2018). In knowledge-grounded dialogue, Ghaddar et al. (2024) show that annotation biases can make models ignore conversation history. Knowledge-gating benchmarks, despite differing in corpora and annotation procedures, have not received comparable scrutiny. Because the benchmarks considered here differ simultaneously in corpus, domain, dialogue format, label distribution, and annotation procedure, our analysis is diagnostic rather than causal: we test whether their labels expose different structural regularities, not whether any single construction factor causes them. This raises two challenges:

Challenge 1 (C1): Do knowledge-gating benchmarks expose exploitable shallow structural regularities? Positive retrieval turns are defined under different corpora and benchmark procedures, including explicit information-seeking annotation, knowledge-grounded enrichment, and subjective knowledge validation. If shallow structural cues reliably predict these labels, benchmark performance may partly rely on regularities that do not require full semantic interpretation.

Challenge 2 (C2): Do such structural rules transfer across benchmark families? Some structural regularities may reflect shared task properties and transfer between related benchmarks, whereas benchmark-specific rules may fail under more dis-

tant transfer. Such failure would limit what can be inferred from strong single-benchmark performance even without identifying the causal source of the mismatch.

We use a deliberately simple probe: logistic regression (LR) over ten dialogue metadata features—position, length, and question-form cues without utterance semantics. If such a probe recovers signal, the benchmark labels contain predictive regularities that do not require full semantic interpretation. This does not imply that semantic retrieval reasoning is absent: the probe is a diagnostic for additional label-predictive structure, not a decomposition of semantic and non-semantic behavior. Across three public task-oriented dialogue benchmarks, we compare feature rankings, permutation tests, and cross-benchmark transfer. Rankings describe structural differences; transfer gives the clearer test. Transfer is strong within DSTC9/DSTC11 but weak across the KETOD–DSTC boundary ($\text{ROC-AUC} = 0.41\text{--}0.48$), and the DSTC→KETOD weakness persists with sentence embeddings and fine-tuned BERT (§4.4).

In summary, we make three contributions:

- We provide evidence that lightweight dialogue-metadata cues expose benchmark-specific regularities in knowledge-gating evaluation.
- We show that strong transfer within the related DSTC9/DSTC11 benchmark family does not imply transfer across the KETOD–DSTC boundary: cross-family ranking quality degrades and positive-class detection becomes weak.
- We show that the observed DSTC→KETOD weakness is not specific to the lightweight structural LR probe: sentence embeddings and fine-tuned BERT exhibit the same qualitative transfer failure.

2 Related Work

Knowledge gating and RAG robustness. Retrieval-augmented generation (Lewis et al., 2020) and knowledge-grounded task-oriented systems (Shi et al., 2023) both face *when* external knowledge is needed. RAGate (Wang et al., 2025) formalizes this decision as binary gating and reports strong in-domain performance in conversational settings. Meaning-preserving query reformulations can also affect retrieval

and downstream generation (Cao et al., 2026). This work studies robustness to input variation within a retrieval system; we instead ask whether gating rules learned from one benchmark remain predictive on another.

Shortcut learning, artifacts, and benchmark validity.

Models can exploit annotation artifacts or shallow correlations rather than intended task features (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019). In knowledge-grounded dialogue, Ghaddar et al. (2024) show that annotation biases in FaithDial can encourage models to ignore conversation history. More broadly, evaluation artifacts can make apparent model weaknesses partly reflect the evaluation procedure, while benchmark construction embeds assumptions about which capabilities are measured and what evidence supports them (Hua et al., 2025; Liu et al., 2024). We follow this diagnostic perspective by testing whether shallow structural cues predict gating labels and whether the resulting rules transfer across benchmarks; feature–label correlation alone is not evidence of a particular causal construction artifact.

Cross-dataset generalization and probing.

Cross-dataset differences can reflect simultaneous changes in data source, task formulation, and annotation, making causal attribution from transfer gaps difficult. This is distinct from data leakage, where overlap can inflate apparent generalization (Elangovan et al., 2021). Our primary structural audit is not a contamination study; for the MiniLM and BERT checks, we separately report exact overlap between source-training and target held-out model inputs after tokenization and truncation to verify that duplicated inputs do not drive the qualitative result. Generalization claims also require attention to transportability across settings (Elangovan et al., 2024). We therefore treat KETOD versus DSTC9/11 as a benchmark-family comparison rather than a controlled decomposition of corpus and annotation effects. Probing classifiers provide a complementary diagnostic of information encoded in models or datasets (Belinkov, 2022); we use a lightweight probe over dataset-level structural attributes.

3 Experimental Setup

3.1 Datasets

Figure 1 summarizes the setup. We evaluate three public task-oriented dialogue benchmarks whose

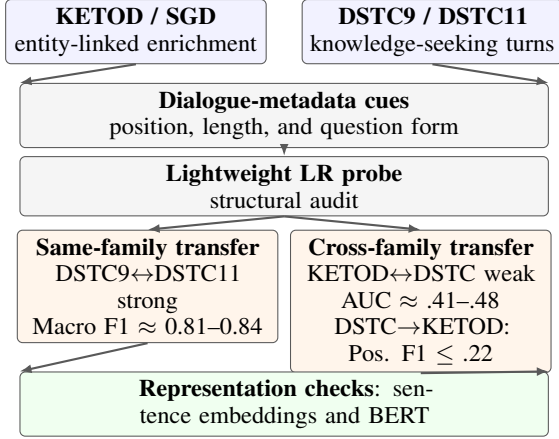


Figure 1: Structural probe and transfer setup. We compare same-family and cross-family transfer, then test whether the weakness persists beyond the lightweight probe.

released annotations can be mapped to a common binary knowledge-gating formulation (Table 1). DSTC9 and DSTC11 form a related MultiWOZ-based pair, while KETOD provides a contrasting SGD-based entity-enrichment setting. KETOD (Chen et al., 2022) enriches SGD dialogues with relevant entity knowledge, including chit-chat-style enrichment; we use *entity-linked enrichment* as a descriptive label rather than an official protocol name. DSTC9 (Kim et al., 2021) extends MultiWOZ 2.1 (Budzianowski et al., 2018) with out-of-API knowledge-seeking turns, while DSTC11 (Kim et al., 2023) focuses on subjective, review-derived knowledge. We use the released validation splits as held-out evaluation sets for DSTC9/11; “test” in our code denotes these held-out splits, not hidden leaderboard tests. We use “KETOD–DSTC” descriptively: because the benchmarks differ along multiple dimensions, transfer gaps should not be interpreted as causal effects of any single construction choice.

Dataset	Train	Held-out	Ratio	Construction
KETOD	41,939	4,964	7.4:1	entity-linked enrichment
DSTC9	71,348	9,663	2.7:1	knowledge-seeking turns
DSTC11	28,431	4,173	1.0:1	subjective knowledge seeking

Table 1: Dataset statistics. Ratio = negative:positive in training; DSTC9/11 held-out evaluation uses the released validation splits.

3.2 Features and Models

We extract ten shallow structural features: positional (2), length-based (5), and question-form (3). They use dialogue metadata and surface indicators such as question marks or question words, but no utterance-level semantics; full definitions are in Appendix A. The goal is not to maximize accuracy, but to test whether minimal structural signals reveal benchmark-associated regularities. We train LR with `class_weight=balanced`, selecting $C \in \{0.01, 0.1, 1.0, 10.0\}$ by stratified cross-validation: 5 folds for Table 2 and 3 folds for Table 3. We also test L1 regularization with degree-2 interactions. Unless stated otherwise, LR uses the default 0.5 threshold, corresponding to source-trained deployment without target-domain calibration. We interpret thresholded F1 with threshold-independent ROC-AUC and average precision (AP; Appendix C). As a robustness analysis (§4.3), we additionally compare a threshold selected from nested source-domain out-of-fold predictions (with C re-selected inside each outer training fold) with an optimistic target-domain threshold-calibration setting, where the classifier remains source-trained but the threshold is selected on the target training split before evaluation on the target held-out split. We report Macro F1 and, for transfer, positive-class F1 and ROC-AUC, since Macro F1 can hide failure on knowledge-needed turns.

4 Results

Sections 4.1–4.3 test shallow structural cues and transfer. §4.4 checks stronger representations, and §4.5 examines the clearest feature-level mechanism.

4.1 Feature Ablation and Permutation

Table 2 compares six feature subsets. On DSTC9/DSTC11, question-form features explain most structural signal: *No Position* matches *Full* (0.80/0.84), *Question Only* reaches 0.77/0.81, and *Position Only* is near-random (0.42/0.33). Shuffling positional features causes $\Delta = 0.000$ on both DSTC datasets, consistent with near-zero coefficients. On KETOD, no group dominates—*Position Only* and *No Position* are similar (0.49 vs. 0.49)—and joint position shuffling drops F1 by 0.064. The datasets therefore differ substantially in which shallow structural cues are label-predictive.

Feature Set	KETOD	DSTC9	DSTC11
Position only	0.489	0.420	0.329
No position	0.487	0.798	0.842
Question only	0.438	0.774	0.809
Length only	0.467	0.547	0.686
Full (10)	0.542	0.798	0.842
L1 + interaction	0.555	0.822	0.847

Table 2: Macro F1 by feature subset. Bold marks the Full setting used in Table 3; L1+interaction is a robustness check.

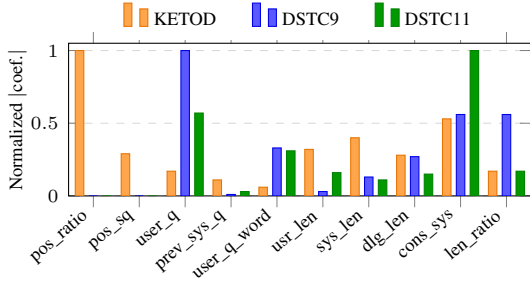


Figure 2: Absolute standardized LR coefficients, normalized within each dataset. Bars show within-dataset relative importance; only ranks enter the Spearman analysis. Zero DSTC position bars reflect constant turn position by task definition.

4.2 Feature Importance: Rankings Diverge

Figure 2 visualizes per-feature importance (absolute standardized LR coefficients). DSTC9 and DSTC11 have nearly identical rankings ($\rho = +0.94$, $p < 0.001$), with `user_has_question` and `consecutive_sys_turns` ranked 1–2. Their correlations with KETOD are negative but not significant ($\rho = -0.27$ and -0.22 ; $p > 0.1$), so we treat ranking divergence as descriptive; transfer metrics and class-conditional rates carry the main evidential weight. Still, the pattern is informative: `consecutive_sys_turns` is consistent (ranks 2/2/1), whereas `turn_position_ratio` is dataset-specific (ranks 1/9/9). The latter is constant on DSTC9/DSTC11 because those benchmarks evaluate the latest user turn, so our claim does not rest on degenerate position features alone.

4.3 Cross-Dataset Transfer

We next test whether structural rules transfer (Table 3). Within the DSTC family, transfer is near-lossless: DSTC9→DSTC11 reaches 0.836 and DSTC11→DSTC9 reaches 0.810, close to or above in-domain scores. Across the KETOD–DSTC boundary, ROC-AUC is weak (0.412–0.481), and DSTC→KETOD positive F1 drops to 0.224/0.209

Train → Held-out	Macro F1	Pos. F1	ROC-AUC
KETOD → KETOD	0.542	0.327	0.719
DSTC9 → KETOD	0.530	0.224	0.481
DSTC11 → KETOD	0.526	0.209	0.474
DSTC9 → DSTC9	0.798	0.740	0.908
DSTC9 → DSTC11	0.836	0.856	0.900
DSTC11 → DSTC9	0.810	0.742	0.906
DSTC11 → DSTC11	0.842	0.852	0.904
KETOD → DSTC9	0.420	0.212	0.463
KETOD → DSTC11	0.427	0.424	0.412

Table 3: Cross-dataset transfer. Positive F1 shows weak DSTC→KETOD detection; ROC-AUC shows weak cross-boundary ordering.

despite Macro F1 being partly masked by KETOD’s majority class. KETOD→DSTC also degrades; KETOD→DSTC11 has non-zero positive F1 but sub-chance ROC-AUC, indicating partly reversed ordering. Thus strong in-domain or same-family transfer does not imply cross-family gating behavior. Full results appear in Appendix C.

Threshold calibration. Beyond the default-0.5 results in Table 3, we compare a threshold selected from nested source-domain out-of-fold predictions with an optimistic target-domain threshold-calibration setting. Target calibration substantially improves positive F1 in the KETOD→DSTC directions (0.047→0.346 and 0.158→0.516), but not consistently for DSTC→KETOD. Because threshold choice cannot affect ROC-AUC, cross-family AUC remains 0.412–0.481, versus about 0.90 within DSTC9/11. Calibration therefore explains some class-specific F1 variation but not the weak cross-family ordering; full results appear in Appendix E.

4.4 Representation Checks: Embeddings and BERT

To test whether the DSTC→KETOD transfer weakness is specific to the lightweight structural representation, we repeat the comparison with MiniLM sentence embeddings plus LR (Reimers and Gurevych, 2019; Wang et al., 2020) and fine-tuned BERT (Devlin et al., 2019). Both use accumulated dialogue context through the evaluated user turn. We explicitly left-truncate at 256 tokens so that the latest turn remains visible and separately check exact input overlap after tokenization and truncation; implementation details are in Appendix B.

Figure 3 shows that fine-tuned BERT drops to

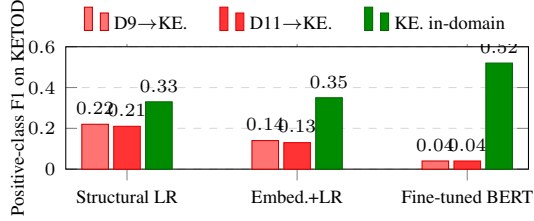


Figure 3: DSTC→KETOD transfer remains weak in the additional representation checks.

Data	P(q pos)	P(q neg)	Δ
KETOD	0.319	0.283	+0.036
DSTC9	0.965	0.312	+0.653
DSTC11	0.968	0.399	+0.569

Table 4: Training-split rate of the current-turn user_has_question feature. The class-conditional gap is large on DSTC9/DSTC11 and small on KETOD.

0.04/0.04 positive F1 in DSTC→KETOD transfer, below embeddings (0.14/0.13) and structural LR (0.22/0.21), despite having the highest KETOD in-domain positive F1 (0.52). Its ROC-AUC is only 0.526/0.498 (AP 0.138/0.130) for DSTC9/DSTC11→KETOD. The MiniLM transfer is weaker still in ranking terms: ROC-AUC is 0.441/0.401 and AP is 0.112/0.104. Thus, the weak cross-family result is not solely a fixed-0.5-threshold effect and is not specific to the lightweight structural representation.

4.5 Case Study: Question Form and Turn Position

Table 4 shows the clearest question-form contrast. user_has_question ranks 7 on KETOD but 1/2 on DSTC9/DSTC11. On DSTC9/11, 96.5%/96.8% of positive examples contain this cue versus 31.2%/39.9% of negatives (gaps of +65.3/+56.9 points); KETOD is nearly flat across labels (31.9% vs. 28.3%; +3.6 points). Thus a DSTC-trained probe that weights this cue has little signal on KETOD. Conversely, turn_position_ratio ranks first on KETOD but last on DSTC, while consecutive_sys_turns ranks 2/2/1; not all structural cues are dataset-specific.

Removing question-form features. To test whether the same-family advantage is driven only by question form, we remove all three question features and retain the two position and five length features, selecting regularization separately under the same 3-fold source-domain CV protocol. DSTC9→DSTC11 and DSTC11→DSTC9 AUC

decrease from 0.900/0.906 to 0.732/0.642, but remain above all four cross-family directions (0.345–0.406). Question form therefore accounts for a substantial part, but not all, of the same-family advantage, while its removal does not eliminate the cross-family separation. Together with calibration, this narrows two specific explanations for the gap without removing it. Full results appear in Appendix D.

5 Conclusion

We audit knowledge-gating benchmarks using shallow structural probes and cross-benchmark transfer. The probes transfer strongly within the related DSTC9/11 family but poorly across the KETOD–DSTC boundary, where ranking quality is weak or partly reversed and DSTC→KETOD positive-class detection is poor. This separation persists under question-form removal and threshold calibration, while sentence embeddings and fine-tuned BERT show the same qualitative DSTC→KETOD weakness. Strong in-domain gating scores therefore should not be interpreted in isolation as evidence of transferable retrieval behavior. We recommend complementing them with cross-benchmark transfer, positive-class metrics, and threshold-independent evaluation.

Limitations

Our probe uses structural features and is intended as a dataset audit rather than a deployable gating model; predictability from shallow features does not imply that semantic retrieval reasoning is absent. Two position features are structurally constant on DSTC9/DSTC11 because those benchmarks evaluate retrieval at the latest user turn by task definition (Kim et al., 2021, 2023). On KETOD, the position features and dialogue_len_log use the completed dialogue’s total USER-turn count and are therefore retrospective rather than online-available. The main transfer conclusions rely on full feature sets, positive-class F1, ROC-AUC, and additional representation checks, so they do not depend on these position features alone.

KETOD and DSTC9/11 differ jointly in corpus, annotation procedure, and label distribution, so the observed transfer gap cannot be attributed causally to any one factor. The question-form, No-Q, and calibration analyses narrow particular explanations but do not remove this confound. DSTC9 and DSTC11 also differ in knowledge type, so treating

them as a DSTC family is an analytical simplification rather than an official taxonomy. Our conclusions are therefore limited to the three benchmarks studied here; a controlled decomposition would require benchmark designs that cross corpus and annotation procedure.

Ethical Considerations

This work uses publicly available benchmark datasets and does not involve human subjects. Our findings are intended to improve evaluation methodology, not to enable misuse. The released reproducibility materials are subject to the licenses of the original datasets and do not redistribute the original benchmark data.

Acknowledgments

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123301), the National Natural Science Foundation of China (No. 62576340), the National Social Science Fund of China (Nos. 22&ZD035 and 25VSZ040), and the Beijing Nova Program (20250484750).

References

- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. [MultiWOZ – a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.
- Tianyu Cao, Neel Bhandari, Akhila Yerukola, Akari Asai, and Maarten Sap. 2026. [Out of style: RAG’s fragility to linguistic variation](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 280–318, Rabat, Morocco. Association for Computational Linguistics.
- Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. 2022. [KETOD: Knowledge-enriched task-oriented dialogue](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, Washington. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Aparna Elangovan, Jiayuan He, Yuan Li, and Karin Verspoor. 2024. [Principles from clinical research for NLP model generalization](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2293–2309, Mexico City, Mexico. Association for Computational Linguistics.
- Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. [Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.
- Abbas Ghaddar, David Alfonso-Hermelo, Philippe Langlais, Mehdi Rezagholizadeh, Boxing Chen, and Prasanna Parthasarathi. 2024. [CHARP: Conversation history AwaReness probing for knowledge-grounded dialogue systems](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1534–1551, Bangkok, Thailand. Association for Computational Linguistics.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. 2025. [Flaw or artifact? rethinking prompt sensitivity in evaluating LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19889–19899, Suzhou, China. Association for Computational Linguistics.
- Seokhwan Kim, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, and Dilek Hakkani-Tur. 2021. [Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access track in DSTC9](#). In *Proceedings of the AAAI-21 DSTC9 Workshop*. DSTC9 Track 1 Overview.
- Seokhwan Kim, Spandana Gella, Chao Zhao, Di Jin, Alexandros Papangelis, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2023. [Task-oriented conversational modeling with subjective knowledge track in DSTC11](#). In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 274–281,

Prague, Czech Republic. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, Virtual.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024. [ECBD: Evidence-centered benchmark design for NLP](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365, Bangkok, Thailand. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tianyuan Shi, Liangzhi Li, Zijian Lin, Tao Yang, Xiaojun Quan, and Qifan Wang. 2023. [Dual-feedback knowledge retrieval for task-oriented dialogue systems](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6566–6580, Singapore. Association for Computational Linguistics.

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](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

Xi Wang, Procheta Sen, Ruizhe Li, and Emine Yilmaz. 2025. [Adaptive retrieval-augmented generation for conversational systems](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 491–503, Albuquerque, New Mexico. Association for Computational Linguistics.

A Feature Definitions

Table 5 lists all ten structural features.

Feature	Definition
<i>Positional</i>	
turn_position_ratio	current USER-turn idx / total USER turns
turn_position_squared	(current USER-turn idx / total USER turns) ²
<i>Length</i>	
user_turn_len_log	log(1+tokens)
sys_turn_len_log	log(1+tokens) for the previous system turn
dialogue_len_log	log(total USER turns)
consecutive_sys_turns	# consecutive sys turns before user
turn_len_ratio	user len / max(sys len, 1), capped at 5
<i>Question</i>	
user_has_question	user turn contains “?”
prev_sys_is_question	prev sys ends with “?”
user_starts_question_word	first token in the fixed question-word list

Table 5: Full feature definitions.

The fixed question-word list is {what, how, where, when, why, is, does, can, do}.

For KETOD, turn_position_ratio, turn_position_squared, and dialogue_len_log use the total number of USER turns in the completed dialogue. They should therefore be interpreted as retrospective metadata for benchmark auditing, not as features available at inference time. For DSTC9/DSTC11, the evaluated user turn is the latest serialized turn, making the two position features constant by task construction.

B Representation-Check Implementations

MiniLM sentence embeddings. We encode the processed accumulated-context input field with all-MiniLM-L6-v2, explicitly setting left truncation and a maximum sequence length of 256. We fit a class-balanced LR on the resulting 384-dimensional embeddings, selecting $C \in \{0.01, 0.1, 1.0, 10.0\}$ by 3-fold source-only cross-validation. Evaluation uses a fixed 0.5 threshold and reports Macro F1, per-class F1, ROC-AUC, and average precision. The random seed is 42.

MiniLM truncation and overlap audits. The tokenizer’s default truncation side is right. On the KETOD/DSTC9/DSTC11 held-out inputs,

20.7%/33.7%/14.2% of examples exceed the 254-content-token capacity. Default right truncation fully retains the final serialized user turn for only 79.3%/66.3%/85.8% of examples, whereas explicit left truncation retains it for 100% in all three splits. Exact input overlap after left truncation is negligible for DSTC9→KETOD (2/4,964) and absent for DSTC11→KETOD. After removing these overlaps, positive F1 remains 0.140/0.135, ROC-AUC is 0.441/0.401, and average precision is 0.112/0.104. Thus, the MiniLM result reflects weak cross-family ranking rather than only a mismatched decision threshold.

Fine-tuned BERT. We fine-tune bert-base-uncased (Devlin et al., 2019) (full model, no layer freezing) for 3 epochs with AdamW (learning rate 2×10^{-5} , weight decay 0.01, linear warmup over 10% of training steps then linear decay), batch size 32, max sequence length 256, and class-frequency-weighted cross-entropy. In all three processed datasets, the input field serializes accumulated dialogue context ending at the evaluated user turn. We left-truncate sequences longer than 256 tokens so that the latest user turn is preserved. We evaluate the final epoch at a fixed 0.5 threshold, without target-domain calibration or development-set early stopping. Random seed is fixed at 42; multi-seed variance is left for future robustness checks.

BERT exact-overlap sensitivity. We report both raw-string overlap and exact input overlap after tokenization and truncation, using the latter as the primary BERT sensitivity. Exact input overlap is negligible for DSTC9→KETOD (2/4,964) and absent for DSTC11→KETOD; removing it leaves positive F1 at 0.038/0.041 and ROC-AUC at 0.526/0.498. Overlap is larger within the related DSTC pair: after removing 910/4,173 exact input overlaps, DSTC9→DSTC11 Macro F1 changes from 0.697 to 0.636 while ROC-AUC remains 0.987; removing 718/9,663 overlaps changes DSTC11→DSTC9 Macro F1 from 0.829 to 0.820 and ROC-AUC to 0.946. Thus, exact input overlap affects some thresholded F1 values but does not account for the weak DSTC→KETOD ranking result.

C Full Transfer Results

Table 6 reports per-class F1, ROC-AUC, and average precision (AP) for all source→held-out combi-

Train → Held-out	Macro	Pos.	Neg.	ROC	AP
KETOD → KETOD	0.542	0.327	0.757	0.719	0.270
KETOD → DSTC9	0.420	0.212	0.628	0.463	0.236
KETOD → DSTC11	0.427	0.424	0.430	0.412	0.440
DSTC9 → KETOD	0.530	0.224	0.836	0.481	0.135
DSTC9 → DSTC9	0.798	0.740	0.857	0.908	0.713
DSTC9 → DSTC11	0.836	0.856	0.816	0.900	0.873
DSTC11 → KETOD	0.526	0.209	0.844	0.474	0.134
DSTC11 → DSTC9	0.810	0.742	0.877	0.906	0.704
DSTC11 → DSTC11	0.842	0.852	0.832	0.904	0.870

Table 6: Full cross-dataset transfer results with per-class F1, ROC-AUC, and average precision (AP). Pos. = positive (knowledge-needed) class, Neg. = negative class. Both transfer directions across the KETOD–DSTC boundary show ROC-AUC near or below 0.5; DSTC→KETOD additionally has positive F1 ≤ 0.22 .

nations.

D Transfer without Question-Form Features

Table 7 compares the full-feature transfer results with models trained after removing all three question-form features. The two position and five length features are retained, and the regularization parameter is selected separately for the full and reduced feature sets under the same 3-fold source-domain cross-validation protocol.

Train → Eval	AUC	Pos. F1	Macro F1
	Full / No-Q	Full / No-Q	Full / No-Q
D9 → D11	0.900/0.732	0.856/0.719	0.836/0.671
D11 → D9	0.906/0.642	0.742/0.486	0.810/0.560
D9 → KETOD	0.481/0.406	0.224/0.179	0.530/0.348
D11 → KETOD	0.474/0.393	0.209/0.182	0.526/0.325
KETOD → D9	0.463/0.397	0.212/0.108	0.420/0.378
KETOD → D11	0.412/0.345	0.424/0.312	0.427/0.377

Table 7: Cross-dataset transfer with the full feature set versus the No-Q variant (D9/D11 = DSTC9/DSTC11). Question form explains a substantial part of same-family transfer, but same-family AUC remains above all four cross-family directions after removal.

E Threshold Calibration Sensitivity

Table 8 compares a threshold selected from nested source-domain out-of-fold predictions with an optimistic target-domain threshold-calibration setting.

Train → Eval	Thr. S / T	Pos. F1 S / T	Macro F1 S / T	AUC
D9 → D11	0.596/0.579	0.844/0.849	0.834/0.837	0.900
D11 → D9	0.500/0.503	0.742/0.743	0.810/0.810	0.906
D9 → KETOD	0.596/0.307	0.217/0.227	0.528/0.530	0.481
D11 → KETOD	0.500/0.669	0.209/0.176	0.526/0.524	0.474
KETOD → D9	0.695/0.377	0.047/0.346	0.409/0.449	0.463
KETOD → D11	0.695/0.404	0.158/0.516	0.374/0.432	0.412

Table 8: Threshold-calibration sensitivity (D9/D11 = DSTC9/DSTC11). S/T gives source-selected/target-selected thresholds; paired F1 values follow the same order. ROC-AUC is threshold-independent.

For the source threshold, each outer fold re-selects C using only its outer-training partition before generating the outer-validation probabilities used for threshold selection. In the target-domain calibration condition, the classifier is still trained only on the source dataset; the target training split is used only to select the decision threshold before evaluation on the target held-out split. Target-domain calibration changes class-specific F1, especially for KETOD→DSTC, but does not change the weak or reversed cross-family ROC-AUC.