

What to Keep in Tight-Budget Dialogue Compression: Turn-Level Information Matters

Minji Park Seunghyun Yoon Hyuk Lim

Korea Institute of Energy Technology (KENTECH)

{minjipark, syoon, hlim}@kentech.ac.kr

Abstract

Dialogue histories often need to be compressed to fit within the context budgets of large language models (LLMs). However, when conversational compression occurs, turns containing critical information are sometimes removed. We study which turns should be retained to minimize information loss during task-oriented dialogue compression. On the MultiWOZ and SGD datasets, we compare four turn-level signals: Frame for task context, Evidence for question-conditioned relevance, Pivot for state changes, and Recency for later context. Evenly spaced selection (Stride) serves as a simple positional baseline.

The effectiveness of each signal depends on the question type. Evidence achieves the highest accuracy among the four signals on current-slot questions in both datasets. At a 30% word budget, Pivot improves over Stride on MultiWOZ when a state change determines the final value (0.421 versus 0.255), but is slightly worse on SGD (0.234 versus 0.245). For questions that require information from both the task context and a final value located in separate turns, Stride performs best among the compressed methods. The best selector depends on the evaluation target, and no single method is best in all settings.

1 Introduction

Task-oriented dialogues often contain a sequence of revisions and decisions. A user may request a restaurant, revise the time, reject an option, and later expect the assistant to continue from the final choice. The *dialogue state* includes currently active slot values, such as restaurant, time, party size, and area. A *pivot turn* changes that state through a revision, replacement, slot update, or task switch. Under a tight memory budget, a compressor may retain the old value but drop the short pivot that determines the current state. The resulting context

T1 System (Frame/Evidence): Golden Wok, a Chinese restaurant, has tables at 7 pm and 8 pm .			
T2 User (Evidence): Book the 7 pm table.			
T3 User (Pivot): Actually, choose the later time .			
Selected context	What the reader sees		Answer
T1, T2, T3 [dropped]	relevant	but stale	7 pm
T1, T2, T3 [kept]	value	plus finalizing update	8 pm

Figure 1: Illustration of late-finalization eviction. Dropping the state-changing turn makes an earlier value appear current.

can remain fluent and relevant while supporting a stale answer.

We define *late-finalization eviction*: compression removes the turn that finalizes the current dialogue state, leaving an earlier state apparently current. The turn containing a value and the turn that makes it final may differ; therefore, retaining an answer string (T2) is not sufficient.

Lexical selection keeps turns that share words with the question; diversity selection keeps varied information to broaden coverage; token compressors preserve tokens that the model considers important; and recency keeps later context. These strategies prioritize different types of information. Lexical selection may retain an answer string without its finalizing update, whereas recency can retain the update but omit earlier supporting context. Diversity-based selection may cover multiple topics without preserving the decisive transition.

We analyze dialogue compression in terms of four turn-level retention signals: Frame, Evidence, Pivot, and Recency. Frame keeps task identity; Evidence keeps turns that are relevant to the question, often including the answer value; Pivot keeps state changes; Recency keeps later local context.

First, Evidence performs best among the four signals for direct current-slot questions (for ex-

ample, “What time is the current reservation?”, “Where is the current destination?”) on both MultiWOZ and SGD datasets. Second, Pivot’s largest gain over Stride is on the MultiWOZ current-slot questions at $r = 0.30$ (+0.166 strict Exact Match, EM); the direction reverses slightly on SGD. Third, at $r = 0.30$, Stride performs best among the compared compressed methods on questions that require both the task context and the final value (0.220 for Stride versus 0.146 for Pivot).

These results suggest that when choosing selectors, we need to consider whether the required evidence is localized or distributed across multiple turns.

The paper makes three contributions.

1. We define four signals for selecting dialogue turns based on task context, relevance to the question, state changes, and turn position.
2. We benchmark the four signals and Stride to examine when signal-based selection improves over simple positional sampling.
3. We distinguish finalizing turns from value-bearing turns and analyze how retaining each relates to answer accuracy. The results show that anchor retention, value visibility, and reader accuracy capture different aspects of compressed dialogue.

2 Related Work

BM25 (Robertson and Zaragoza, 2009), MMR (Carbonell and Goldstein, 1998), and prompt compressors such as LLMLingua-2 (Pan et al., 2024) and LongLLMLingua (Jiang et al., 2024) select or preserve information based on relevance, diversity, or token importance. Dialogue state tracking models instead track the user’s current state over a conversation, including revisions and previously unseen values (Henderson et al., 2014; Wu et al., 2019). We use this state information to decide which dialogue turns to retain during compression. A revision can make an earlier value outdated, so retaining the turn where the value changes can be important for recovering the correct answer.

3 Method

Diagnostic Signals Let a dialogue be $D = (t_0, \dots, t_{n-1})$ and q a query. A compressor selects turns under a word budget $B = \lfloor r \sum_i c_i \rfloor$, where

c_i is the number of words in turn t_i . Selected turns are restored to dialogue order and passed to a frozen reader. Each turn receives

$$S_i = [F_i, E_i, P_i, R_i]. \quad (1)$$

The four scores are used alone or in a uniform mixture; turns are ranked by score per word cost and greedily selected within B .

Answerability may depend on multiple turns, particularly when a value is introduced in one turn and later revised in another. We therefore measure both downstream reader accuracy and whether the selected context kept the evidence that makes the answer current.

Signal		Role
Frame	F	Task, domain, and intent cues, often near the opening turns.
Evidence	E	Relevance of the turn to the question.
Pivot	P	Revisions, replacements, slot updates, switches.
Recency	R	Turn position, favoring later context.

Table 1: Four retention signals.

Frame gives higher scores to earlier turns using the positional prior $2^{-i/2}$ for turn index i . It also assigns a separate score based on task- and domain-related cues in the turn. The final Frame score is the larger of these two values. Evidence scores each question-turn pair with the frozen cross-encoder, which is based on M3 (Chen et al., 2024). The cross-encoder outputs a relevance score, which is converted to a value between 0 and 1 using a sigmoid function. Thus, Evidence measures how relevant each turn is to the question rather than simply counting shared words. Recency uses $R_i = i/(n - 1)$ (or 1 for a single-turn dialogue), so later turns receive higher scores regardless of their content. Implementation details are given in Appendix A.6.

In Figure 1, Evidence can prefer the turns containing the two times, whereas Pivot can prefer the short phrase “choose the later time” even though it does not repeat the final value “8 pm.” The uniform-mixture control gives each signal weight 0.25. It tests whether averaging different scopes is better than choosing the scope matched to the evaluation target.

Pivot scoring and supervision. A turn is labeled as a Pivot (positive) when the tracked state differs from the preceding state on any slot; this

difference is the state delta used as its training label. A previous/current-turn pair is encoded by a frozen BERT encoder (Devlin et al., 2019); only a two-layer multilayer perceptron (MLP) is fit on training-split state deltas and then frozen. Pivot receives training labels. However, Stride, BM25, MMR, and the prompt-compression baselines do not receive Pivot labels. Accordingly, Pivot–Stride gaps reflect both the Pivot signal and task-specific supervision. No state labels or reader outputs are required at inference; the already trained scorer reads each adjacent turn pair and assigns its Pivot score.

4 Experimental Setup

We use evaluation examples from MultiWOZ 2.2 (Budzianowski et al., 2018; Zang et al., 2020) and Schema-Guided Dialogue (SGD) (Rastogi et al., 2020). This analysis needs a turn-level reference to the update that makes a value current. The dialogue-state annotations in both corpora allow us to construct deterministic finalizing anchors without additional human annotation. Their different schemas and collection procedures provide a cross-corpus check on the selector rankings.

At each user turn, we compare the current state with the preceding state and record every slot whose normalized value changes. We require the normalized old and final values to occur somewhere in the full dialogue, then retain the last update to each candidate slot. The aligned user turn defines the single finalizing anchor $A(p)$. The surface-value requirement places updates whose final value remains entirely implicit outside the evaluation scope.

Canonical slot identifiers are lowercased, with hyphens and underscores rendered as spaces. The CURRENT-SLOT question inserts this readable name into a fixed request for its current value; Appendix A.6 specifies the reference point, requested fields, and output format. For TASK-AND-CURRENT-SLOT, the opening goal reference is the first sentence of the opening user turn. This span serves as a deterministic proxy for the opening goal, and content-word scoring permits paraphrased task descriptions. Joint examples have at least ten turns; the final update occurs at index six or later and at or beyond the dialogue midpoint. We also remove examples whose gold value is already visible in the 30%-budget Prefix context.

The primary budget is $r = 0.30$. It means that a

selector may retain at most 30% of the dialogue’s whitespace-word count. Appendix A.1 covers $r \in \{0.10, 0.30, 0.50, 0.70\}$. Our main analyses focus on CURRENT-SLOT and TASK-AND-CURRENT-SLOT. We therefore exclude standalone TASK results from the analyses below. MultiWOZ contributes 140 and SGD 500 examples per seed and target for seeds $\{42, 43, 44\}$. The smaller filtered MultiWOZ pool sets its cell size; the larger SGD pool is capped at 500 to keep the reader sweep tractable. Each seed resamples probes without replacement. Reader decoding is deterministic, and every compressor receives the same sampled cell. The two target pools can share a dialogue–slot instance: across seeds, the overlap is 110–121 instances for MultiWOZ and 354–375 for SGD. When combining results across the two datasets, we weight each dataset by its number of examples rather than averaging the two dataset scores equally.

Readers, metrics, and controls. The primary reader is Llama-3.1-8B-Instruct (Grattafiori et al., 2024). We evaluate Pivot-only with Qwen2.5-7B-Instruct (Yang et al., 2024) to test whether its budget trend transfers to a second reader family. Greedy decoding uses 128 new tokens. The reader is frozen and does not participate in turn selection. We use different evaluation rules for the two targets. For current-slot, prediction and gold are normalized with lowercasing, punctuation, whitespace, and time normalization (e.g. 7 pm is equivalent to 19:00). We also treat different written forms of the same MultiWOZ value as equivalent under the fixed normalization rules listed in Appendix A.6.

Strict Exact Match (Strict EM) requires the normalized prediction to be identical to the normalized gold answer. Loose EM is more permissive. It also accepts cases in which one normalized string contains the other. The TASK-AND-CURRENT-SLOT target requires two pieces of information: the task and the final slot value. We therefore score these two parts separately instead of comparing the full generated response as one string.

For the task part, the prediction must share at least one meaningful content word with the gold task description. Common function words are ignored. For the value part, we report two scores.

Our main score is *value-string composite accuracy*. A prediction is counted as correct only if it

satisfies the task condition and contains the complete normalized gold value.

We also report a more permissive measure, *value-overlap composite accuracy*. This score allows partial overlap with the gold value. The required overlap is $\max(1, \lfloor |\text{gold content tokens}|/2 \rfloor)$ content tokens. We use value-string composite accuracy for the main results and report value-overlap accuracy as a secondary measure in the appendix. Additional prompt and normalization details are given in Appendix A.6.

We also use two simple diagnostics to check what information survives compression. *Anchor retention* asks whether the turn that finalizes the queried value is still present in the compressed dialogue. *Gold-value visibility* asks whether the accepted answer string appears anywhere in the compressed dialogue.

A compressed dialogue can keep the finalizing turn but still omit information needed to recover the answer. It can also keep the answer string while dropping the update that makes that value current.

We compare the four individual signals and their uniform mixture with several baselines. BM25 selects turns based on lexical relevance (Robertson and Zaragoza, 2009). MMR also considers relevance, but tries to reduce redundancy among the selected turns (Carbonell and Goldstein, 1998).

LLMLingua-2 (Pan et al., 2024) and LongLLM-Lingua (Jiang et al., 2024) work at the token level rather than selecting complete turns.

We also include position-based controls. Prefix keeps the earliest turns, Tail keeps the latest turns, Stride spreads selections across the dialogue, and Random samples turns with a fixed seed. Full Context is the uncompressed reference. Tables A.2 and A.3 report the dataset-wise results for these controls.

Leakage guard. To avoid leakage from the evaluation set, none of the selection methods uses held-out answers, anchor annotations, gold slot values, or reader outputs. Pivot is trained only from training-split state changes and then frozen. Evidence scores are computed from the question and candidate turn, not from the gold answer. The artifact checks that dialogues used to train Pivot do not overlap the held-out probe split.

5 Results

Method	SGD	MultiWOZ
Stride	0.245±.016	0.255±.011
Recency-only	0.180±.012	0.398±.030
Frame-only	0.127±.001	0.200±.019
Evidence-only	0.370±.014	0.509±.025
Pivot-only	0.234±.019	0.421±.007
Uniform mixture	0.157±.014	0.348±.030

Table 2: CURRENT-SLOT strict EM at $r = 0.30$ (mean±standard deviation across three seeds).

The Pivot-Stride comparison changes by dataset. On MultiWOZ current-slot, Pivot reaches 0.421 strict EM compared with 0.255 for Stride. On SGD, however, Pivot is slightly lower than Stride (0.234 vs. 0.245). Thus, Pivot’s advantage depends on the dataset.

Evidence leads the four primary signals for direct slot recovery. Among the four primary signals, Evidence-only gives the highest current-slot strict EM on both datasets: 0.370 on SGD and 0.509 on MultiWOZ (Table 2). Evidence targets turns relevant to the question, and Pivot scores state changes unrelated to the queried slot.

Method	Value-string comp. acc.
Pivot-only	0.146
Uniform mixture	0.122
BM25	0.099
MMR	0.122
LLMLingua-2	0.125
LongLLMLingua	0.062
Stride	0.220
Full Context	0.580

Table 3: Pooled TASK-AND-CURRENT-SLOT value-string composite accuracy at $r = 0.30$ ($n = 1,920$).

For the joint target, the reader must recover both task identity and the final slot value, which may be supported by separated turns. At $r = 0.30$, Stride reaches 0.220 value-string composite accuracy, exceeding every compressed row shown in Table 3; Pivot reaches 0.146. The dataset-wise appendix shows the same Pivot-Stride ordering on both datasets. Pivot may prioritize detected state changes, while Stride spreads its selections across the dialogue.

Full-context reference. Full Context selects every dialogue turn and reaches 0.580 value-string composite accuracy on TASK-AND-CURRENT-SLOT. On CURRENT-SLOT, its strict EM is 0.659

on SGD and 0.676 on MultiWOZ. Thus, some errors persist even without compression. Because all selectors are evaluated on the same sampled examples with the same reader, prompt, and scoring procedure, their relative differences are still comparable.

6 Discussion

For probe p , let $A(p)$ be the annotated turn where the queried value becomes final, $V(p)$ the turns containing an accepted answer string, and $C(p)$ every turn where any tracked slot changes. These sets need not coincide. In Figure 1, T3 is the finalizing anchor, T2 contains the old value, and deriving “8 pm” requires combining the candidate times in T1 with the update in T3. More generally, reader answerability can require information from $A(p) \cup V(p)$, while task identity can require an additional early frame turn.

Evidence favors query-relevant turns and may therefore overlap with $V(p)$, whereas Pivot targets the broader set $C(p)$. Because $C(p)$ can also include changes to slots unrelated to the question, Pivot may spend part of the budget on genuine but irrelevant updates. Recency can retain a late $A(p)$ while dropping an earlier turn that contains the value or task context. This is why anchor retention, value visibility, and reader accuracy do not always rank the methods in the same order.

Keeping the finalizing turn alone does not guarantee answer recovery. At the same time, keeping the answer string alone does not mean that value is still current. Evidence is more useful when the required information is concentrated in a few query-relevant turns. Pivot is more useful when relevant revisions are frequent. Stride is better suited where the required evidence is spread across distant turns.

7 Conclusion

The best retention signal depends on the evaluation target. Evidence is strongest among the four primary signals for direct current-slot recovery on both datasets. Pivot has a clear supervised advantage over positional coverage specifically on MultiWOZ current-slot, but is slightly worse on SGD. Stride is stronger on pooled joint composite accuracy, where the reader needs broad context as well as the final state.

Limitations

MultiWOZ and SGD are English, task-oriented, schema- and slot-based datasets. Their explicit states make revision structure measurable. The results do not establish that these findings generalize to open-ended or multi-party dialogue, long-form assistant interactions, preferences, rationales, safety constraints, or multi-turn reasoning chains. Future work should test robustness to prompt variation and broader dialogue types, and measure end-to-end latency.

Acknowledgments

This work was partially supported by the Korea Planning & Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIR) (Alchemist Project RS-2025-02634277) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (RS-2021-II210379, Artificial Intelligence Innovation Human Resources Development RS-2026-25548323, and AI Star Fellowship Support Program RS-2026-25614738), and the Korea Innovation Foundation (INNOPOLIS) grant funded by the Korea government (MSIT) (RS-2026-25549080).

Claude and Codex were used for language editing and for feedback on the clarity and presentation of manuscript drafts. The authors reviewed all suggestions and take full responsibility for the final manuscript.

Ethical Considerations

We use released task-oriented dialogue datasets and do not collect new data from humans. Compression may drop an important user constraint or update even when the remaining context looks reasonable.

References

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.

- Jaime G. Carbonell and Jade Goldstein. 1998. [The use of MMR, diversity-based reranking for reordering documents and producing summaries](#). In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336. ACM.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. 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.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. [The second dialog state tracking challenge](#). In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 263–272. Association for Computational Linguistics.
- Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. [LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677. Association for Computational Linguistics.
- Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. 2024. [LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 963–981, Bangkok, Thailand. Association for Computational Linguistics.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. [Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. [Transferable multi-domain state generator for task-oriented dialogue systems](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. [MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines](#). In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Appendix

A.1 Reader Accuracy Ratio Sweep

Value-string composite accuracy				
Method	.10	.30	.50	.70
Stride	.003	.220	.308	.444
Frame-only	.024	.090	.217	.376
Evidence-only	.017	.130	.274	.381
Pivot-only	.025	.146	.269	.350
Recency-only	.010	.045	.118	.266
Uniform mixture	.020	.122	.228	.315
BM25	.019	.099	.230	.372
MMR	.025	.122	.238	.367
LLMLingua-2	.014	.125	.282	.461
LongLLMLingua	.021	.062	.124	.176

Value-overlap composite accuracy				
Method	.10	.30	.50	.70
Stride	.041	.353	.474	.591
Frame-only	.067	.245	.423	.546
Evidence-only	.041	.219	.369	.491
Pivot-only	.052	.293	.475	.559
Recency-only	.016	.076	.182	.328
Uniform mixture	.040	.223	.408	.519
BM25	.051	.217	.368	.504
MMR	.048	.199	.337	.464
LLMLingua-2	.065	.217	.359	.543
LongLLMLingua	.048	.129	.222	.284

Table A.1: Pooled TASK-AND-CURRENT-SLOT composite accuracy across ratios, using the Llama reader and three seeds.

At $r = 0.10$, most contexts lack a value-bearing turn, finalizing pivot, or both. Differences emerge at $r = 0.30-0.50$ and remain method-dependent at $r = 0.70$: LLMLingua-2 has the highest value-string accuracy there, whereas Stride has the highest value-overlap accuracy.

A.2 Dataset-Wise Primary Results

Method	SGD		MultiWOZ	
	Overlap	String	Overlap	String
Prefix	.182±.015	.011±.006	.095±.004	.010±.008
Tail	.090±.007	.072±.007	.033±.011	.024±.008
Stride	.358±.034	.215±.019	.333±.018	.238±.027
Random	.217±.024	.135±.019	.233±.054	.193±.052
Full Context	.677±.044	.582±.049	.648±.033	.574±.022
Recency-only	.065±.003	.033±.005	.114±.038	.086±.037
Frame-only	.248±.038	.081±.015	.233±.023	.121±.025
Evidence-only	.227±.023	.128±.015	.190±.018	.138±.011
Pivot-only	.287±.046	.124±.025	.312±.015	.226±.008
Uniform mixture	.200±.039	.093±.026	.305±.027	.226±.029

Table A.2: TASK-AND-CURRENT-SLOT composite accuracy at $r = 0.30$ by dataset (mean±standard deviation across seeds).

Method	SGD	MultiWOZ
Prefix	.015±.001	.009±.004
Tail	.247±.023	.202±.004
Stride	.245±.016	.255±.011
Random	.266±.014	.317±.032
Full Context	.659±.013	.676±.043

Table A.3: Current-slot strict EM for positional controls and Full Context at $r = 0.30$ (mean±standard deviation across seeds).

For the joint target, Stride exceeds Pivot on both datasets. On current-slot, Random exceeds Stride on SGD and MultiWOZ; Prefix and Tail remain lower than both controls.

A.3 Anchor Retention and Hybrid Controls

Method	$r = .10$	.30	.50	.70
Recency-only	.400	.834	.970	.997
MMR	.165	.420	.614	.777
BM25	.189	.385	.547	.685
LLMLingua-2	.000	.273	.836	.992
LongLLMLingua	.041	.073	.197	.354

Table A.4: Late-finalization anchor retention on joint probes ($n = 1,920$ per cell).

Method	Anchor retained	Gold value visible	Reader accuracy
Frame-only	.219	.155	.090
Evidence-only	.356	.454	.130
Pivot-only	.759	.303	.146
Recency-only	.834	.291	.045
Uniform mixture	.734	.231	.122
Prefix	.000	.000	.011
Tail	.854	.369	.061
Stride	.124	.336	.220

Table A.5: Information retained by compressed joint contexts at $r = .30$, pooled over both datasets and three seeds ($n = 1,920$). Reader accuracy is value-string composite accuracy. Bold marks the highest compressed score in each column.

Tail retains the most anchors, Evidence preserves the gold value most often, and Stride gives the highest reader accuracy. The three measures therefore capture different properties of the compressed context.

$r = .10$				
Method	MW	SGD	All	Δ All
BM25+Pivot	.333	.407	.391	+.202
MMR+Pivot	.291	.345	.333	+.168

$r = .30$				
Method	MW	SGD	All	Δ All
BM25+Pivot	.662	.708	.698	+.313
MMR+Pivot	.581	.706	.679	+.259

Table A.6: Anchor retention for hybrid selectors across three seeds. Δ All is the gain over the corresponding BM25 or MMR row in Table A.4.

Adding Pivot raises anchor retention for BM25 and MMR at both budgets. Reader accuracy still depends on the other information retained in the context, as Table A.5 illustrates.

A.4 Reader Portability

Value-string composite accuracy				
Reader	.10	.30	.50	.70
Llama-3.1-8B	.025	.146	.269	.350
Qwen2.5-7B	.023	.134	.239	.322

Value-overlap composite accuracy				
Reader	.10	.30	.50	.70
Llama-3.1-8B	.052	.293	.475	.559
Qwen2.5-7B	.049	.281	.437	.519

Table A.7: Pooled Pivot-only portability across readers and ratios.

A.5 Pivot Scorer Constructions

Table A.8 changes only the Pivot score; the probes, selector, reader, and evaluation remain fixed. GoldDelta assigns 1 when a turn changes any tracked slot and 0 otherwise. Cue/value adds 0.65 for a revision or status expression and 0.25 for an explicit time, date, price, area, or Boolean value. On turns after the first, the correction terms *no*, *not*, *instead*, *actually*, and *change* add 0.20. Scores are clipped at 1. All cells use Llama-3.1-8B-Instruct, $r = 0.30$, and seeds 42-44.

GoldDelta improves over learned Pivot on both datasets but remains below Evidence-only. Cue/value trails learned Pivot on MultiWOZ by 0.047 and exceeds it on SGD by 0.236.

A.6 Key Implementation Settings

We did not perform a hyperparameter search; we used the fixed settings reported in Table A.9.

Pivot scorer	MultiWOZ	SGD
Learned	.421 $\pm$.007	.234 $\pm$.019
GoldDelta	.457 $\pm$.014	.332 $\pm$.015
Cue/value	.374 $\pm$.033	.470 $\pm$.009

Table A.8: Current-slot strict EM for Pivot scorer constructions (mean $\pm$ standard deviation). GoldDelta uses evaluation-time states; Cue/value is annotation-free but mixes change and value-surface cues.

Component	Setting
Pivot	Frozen bert-base-uncased; two-layer MLP; 2 epochs; AdamW, 3×10^{-5} ; maximum pair length 256.
Frame	Fixed intent/domain cues combined with the early-position prior.
Evidence	Frozen BAAI/bge-reranker-v2-m3; sigmoid-normalized score.
BM25	Lowercased word tokens; $k_1 = 1.2$, $b = 0.75$; score per word.
MMR	all-MiniLM-L6-v2; $\lambda = 0.5$.
Token compressors	LLMLingua-2 and LongLLMLingua checkpoints; outputs capped at the whitespace-word ceiling B .

Table A.9: Key implementation settings.

A.7 Prompt and Scoring Details

Prompt structure and task scoring. The reader is restricted to the supplied compressed context and returns one unfenced JSON object. CURRENT-SLOT requests one short value after the latest visible update. TASK-AND-CURRENT-SLOT uses the layout `<task>; {slot}: <final value>`. Both prompts include `value`, `support`, and `abstain`; support is capped at 30 words.

For the joint target, the gold task phrase is lowercased, stripped of trailing punctuation and common request prefixes, and tokenized with `[a-z0-9]+`. We remove short tokens and a fixed English stopword list. The task component is correct when prediction and reference share a remaining token.

Normalization. We apply Unicode NFKC, lowercase text, canonicalize time expressions to a zero-padded 24-hour form, remove punctuation other than colons, slashes, and hyphens, and collapse whitespace. Strict EM tests equality after normalization; Loose EM also accepts containment in either direction.