Journal of the Korean Institute of Industrial Engineers
Vol. 52, No. 2, pp. 116-131, April 2026.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2026.52.2.116

LLM 기반 요약 유용성 예측을 활용한 다중 세션 대화 시스템 메모리 관리 효율화

이수연 · 조성준†

서울대학교 산업공학과

Efficient Memory Management in Multi-Session Dialogue Systems via LLM-Guided Summary Usefulness Prediction

Suyeon Lee · Sungzoon Cho

Department of Industrial Engineering, Seoul National University

Long-term dialogue systems often suffer from degraded memory quality as accumulated session summaries introduce redundancy, noise, and hallucinated content. Prior work has mainly focused on improving summarization quality, leaving the usefulness of individual summary items largely unexplored. This paper presents a memory management framework that predicts the usefulness of each summary item using pseudo labels generated through an LLM-as-a-Judge procedure and prunes low-value content before it enters the memory. The proposed model follows a two-stage training strategy: it first learns a usefulness prediction module on top of a frozen encoder, then applies LoRA-based adaptation to refine encoder representations for task-specific discrimination. Experiments show that the model more accurately captures the relative importance of summary bullets, and dialogues conditioned on pruned memory exhibit more coherent and contextually appropriate responses. These findings highlight that managing the quality of stored memory, rather than expanding its volume, is crucial for improving long-term dialogue performance.

Keywords: Multi-Session Dialogue System, Dialogue Memory, Large Language Models

1. 서 론

대화 시스템 연구는 초기의 규칙 기반 접근법에서 시퀀스-투-시퀀스(Seq2Seq) 모델을 거쳐 트랜스포머(Transformer) 기반 구조로 발전함에 따라 점차 더 복잡하고 자연스러운 응답 생성이 가능해졌다(Yi *et al.*, 2024). 최근에는 대규모 언어 모델(Large Language Models; LLM)의 급격한 발전으로 인간 수준에 근접한 자연스러운 대화와 상황 인식 기반의 맞춤형 응답 생성이 가능해지면서, 대화 품질 전반에서 획기적인 향상이 이루어지고 있다(Wang *et al.*, 2023).

이러한 발전으로 단일 세션 내의 문맥 이해를 넘어 장기적 상호작용과 사용자 특성을 반영하는 개인화(personalization) 대화 시스템에 대한 관심이 높아지고 있다. 과거 대화 발화, 사용자 페르소나, 선호도 정보 등을 저장하고 활용하는 검색 증강 생성(Retrieval-Augmented Generation; RAG) 기반 접근이 활발히 연구되고 있으며(Kasahara *et al.*, 2022; Zhong *et al.*, 2022; Jang *et al.*, 2023; Chen *et al.*, 2024), 그중에서도 여러 세션에 걸쳐 축적되는 사용자 정보를 활용하는 메모리 기반 대화 시스템(memory-augmented dialogue systems)이 특히 주목받고 있다(Jo *et al.*, 2024; Tan *et al.*, 2025).

메모리 기반 대화 시스템에서는 일반적으로 세션 종료 시 대화 내용을 요약하여 장기 메모리로 누적하는 방식이 널리 사용된다(Wang *et al.*, 2025). 특히 최근에는 LLM을 활용해 대화 세션을 자동으로 요약하고 이를 장기 메모리로 저장하는

† 연락저자 : 조성준 교수, 08826 서울시 관악구 관악로1 서울대학교 공과대학 산업공학과, Tel : 02-880-7025, Fax : 02-889-8560, E-mail : zoon@snu.ac.kr
2025년 12월 5일 접수; 2026년 1월 1일 수정본 접수; 2026년 1월 12일 게재 확정.

접근이 활발히 연구되고 있다(Wang *et al.*, 2025; Liu *et al.*, 2025). 그러나 생성된 요약 정보를 별도의 관리 없이 그대로 저장할 경우, 메모리 규모가 빠르게 증가할 뿐 아니라 중복되거나 부정확한 정보, 의미적 중요도가 낮은 항목까지 함께 축적되어 검색 효율 저하 및 응답 생성 품질 저하로 이어질 수 있다. 기존 연구는 주로 요약 자체의 품질 향상이나 요약 기법 자체의 개선에 초점을 맞추어 왔으며, 세션 요약 내 개별 항목 중 어떤 정보가 이후 대화에서 실제로 유용한지를 정량적으로 평가하고 관리하는 문제는 상대적으로 충분히 다루어지지 않았다(Wang *et al.*, 2025; Tan *et al.*, 2025). 이는 실제 메모리 기반 대화 시스템에서 중요하지 않은 정보가 과도하게 누적되거나, 반대로 핵심 정보가 적절히 활용되지 못하는 문제로 이어질 수 있다(Pan *et al.*, 2025).

이러한 한계를 해결하기 위해 본 연구는 LLM이 생성한 세션 요약의 개별 항목을 대상으로, 이후 대화에서의 활용 가능성을 기준으로 유용성(usefulness)을 정량화하고 이를 기반으로 메모리를 선택적으로 유지하고 제거하는 요약 기반 메모리 관리 프레임워크를 제안한다. 본 연구에서 유용성이란 특정 요약 항목이 다음 세션의 실제 발화 생성에 기여하는 정도를 의미하며, 이는 요약 메모리가 실제 대화 생성 단계에서 활용되는 사용 시나리오를 직접적으로 반영한다. 또한 메모리 관리에 특화된 학습 데이터가 부족하다는 현실적 제약을 고려하여, LLM-as-a-Judge 방식을 통해 pseudo label을 구축하고, 사전 학습 인코더를 기반으로 한 약지도(weak supervision) 환경에서 2단계 경량 학습 전략을 설계하였다.

실험 결과, 제안한 프레임워크는 회귀 및 랭킹 기반 평가 지표 전반에서 일관된 성능 향상을 보였으며, 메모리 pruning을 통해 입력 메모리 규모와 토큰 수를 크게 감소시키는 동시에 대화 응답 품질을 개선하였다. 이러한 결과는 본 연구의 접근이 대규모 메모리 환경에서도 확장 가능한 개인화 장기 대화 시스템 설계에 효과적으로 기여할 수 있음을 보여준다.

2. 관련 연구

2.1 LLM을 활용한 대화 시스템

대규모 언어 모델(LLM)의 급속한 발전은 대화 시스템 연구의 흐름과 방향을 크게 바꾸어 놓았다. 기존의 트랜스포머 기반 언어 모델이 주로 문장 단위의 맥락 처리에 초점을 맞추었던 데 반해, LLM은 대규모 사전학습을 통해 장문의 입력에서도 높은 표현력을 발휘한다(Brown *et al.*, 2020; Touvron, 2023). 또한, 다양한 도메인에 대한 일반화 능력과 복잡한 의도 추론 능력을 갖추어 인간과 유사한 수준의 자연스러운 발화를 생성할 수 있게 되었다(Wang *et al.*, 2023). 이로 인해 LLM 기반 대화 생성은 현재 대화 시스템 연구의 주류 접근으로 자리 잡고 있다.

또한 Instruction tuning(Ouyang *et al.*, 2022), Preference optimization(Rafailov *et al.*, 2023), 프롬프트를 활용한 역할 및 과정 지시(Ouyang *et al.*, 2022; Wei *et al.*, 2022) 등 LLM을 조정하는 기법이 발전하면서, 최근 연구들은 LLM을 단순한 응답 생성 모델을 넘어 대화 평가자(judge), 요약기(summarizer), 기억 생성기(memory writer) 등 다양한 기능을 수행하는 구성 요소로 활용하고 있다(Lu *et al.*, 2023; Zhong *et al.*, 2024; Wang *et al.*, 2025; Tan *et al.*, 2025).

그러나 LLM만을 단독으로 사용하여 장기 대화를 유지하는 데에는 구조적 제약이 존재한다. LLM은 고정된 입력 창(context window)에 의존하기 때문에 장기간의 대화를 그대로 입력으로 유지하기 어렵고, 장기적 맥락(long-term dependency)을 지속적으로 추적하고 관리하는 능력에도 한계가 있다(Maharana *et al.*, 2024; Wang *et al.*, 2025; Chhikara *et al.*, 2025). 이러한 이유로 장기 맥락을 별도 모듈로 저장하고 필요한 경우 검색해 LLM에 제공하는 메모리 기반 대화 시스템이 활발히 연구되고 있다(Xu *et al.*, 2022; Zhong *et al.*, 2024; Li *et al.*, 2025b).

2.2 메모리 시스템 기반 장기 대화 시스템

장기 대화 시스템의 목표는 단일 세션을 넘어 지속적인 상호작용에서 사용자 정보, 대화 이력, 선호도, 맥락적 사건 등을 기억하여 응답에 반영하는 것이다. 이를 위해 다양한 메모리 구조가 제안되어 왔으며, 대표적으로 대화 내역 전체를 저장하는 방식(Zhong *et al.*, 2024), 세션 단위 요약을 기반으로 메모리를 압축하는 방식(Kang *et al.*, 2025; Wang *et al.*, 2025), 페르소나와 같은 사용자 특성이나 과거 이벤트 등을 메모리로 관리하는 방식(Xu *et al.*, 2022; Zhong *et al.*, 2024) 등이 있다. 이러한 메모리는 일반적으로 검색(retrieval) 모듈과 결합하여 대화의 일관성과 개인화를 강화하며, 사용자 참여도를 높이는 역할을 수행한다.

예를 들어, MemoChat(Lu *et al.*, 2023)은 LLM이 스스로 메모를 생성해 업데이트하도록 설계된 구조로, 대화 내용을 토픽별로 분류하고 장기적 관련성이 높은 정보를 지속적으로 유지하도록 한다. MemoryBank(Zhong *et al.*, 2024)는 대화 기록, 사용자 이벤트, 페르소나와 같은 사용자 특성을 각각 요약해 계층적 메모리 구조로 저장하며, 시간이 지남에 따라 중요도가 감소하는 정보를 망각곡선 기반 감쇠(decay)로 관리하는 LLM에 최적화된 메모리 메커니즘을 제안했다.

RecurSum(Wang *et al.*, 2025)은 LLM이 세션 요약을 재귀적으로 생성하고 업데이트함으로써 글로벌 요약을 유지하는 방법을 제안하였다. Kang *et al.*(2025)은 기존 메모리 연구(Bae *et al.*, 2022)에서 사용되어 온 DELETE, REPLACE 등의 메모리 관리 연산의 한계를 지적하고, 이전 세션 요약과 현재 세션 요약을 결합하여 메모리를 동적으로 재구성하는 접근법을 제안하였다. 또한 PLATO-LTM(Xu *et al.*, 2022)과 LD-Agent(Li *et*

al., 2025)는 페르소나 및 사용자 이벤트 정보를 장기 메모리에 축적하고 검색하여 활용함으로써 대화의 장기적 일관성을 강화하였다. THEANINE(Ong *et al.*, 2025)은 메모리 간의 시간 및 인과적 관계를 기반으로 메모리를 연결하여 타임라인 형태의 구조를 구성하는 메모리 관리 방식을 제안하였다.

이들 연구는 장기 대화를 위해 메모리를 어떻게 저장하고 검색할 것인지에 주로 초점을 두어 메모리 구조와 요약 생성 방법을 고도화해 왔으며, 해당 영역에서 다양한 접근이 지속적으로 제안되어 왔다. 그러나 저장된 요약 중 어떤 정보가 실제로 이후 대화에 유용한지를 판단하기 위한 정량적 기준을 학습 기반으로 제시한 연구는 매우 제한적이다. 본 연구는 이러한 공백을 메우기 위해, 메모리 항목 단위의 유용성을 평가하고 중요도가 낮은 항목을 자동으로 제거하는 메모리 관리 모델을 제안한다는 점에서 기존 접근과 차별성을 가진다.

3. 방법론

본 연구의 목표는 장기 대화 시스템에서 세션 간 맥락을 유지하기 위해 활용되는 요약 기반 메모리(summary memory)의 품질을 향상시키는 데 있다. 이를 위해 각 세션 종료 시 LLM을 활용하여 세션 요약을 생성하고, 별도의 경량 메모리 관리 모델을 통해 각 요약 항목의 유용성을 예측하여 중요도가 낮은 요약 항목을 제거(pruning)하는 프레임워크를 제안한다. 전체 구조는 <Figure 1>에 제시되어 있다.

3.1 세션 단위 메모리 생성

각 대화 세션이 종료되면 LLM을 활용해 해당 세션의 대화 내용을 bullet point 형태로 요약한다. 이때 단순 요약 생성에 그치지 않고, 각 bullet이 어떤 발화를 근거로 도출되었는지를 함께 명시하도록 프롬프트를 설계하여 요약의 정합성을 높였다. 또한 생성된 요약이 이후 응답 생성 과정에서 활용된다는 점을 LLM에 명확히 안내함으로써 목적 적합성을 강화하였다.

각 대화 세션 $S_t = \{u_1,...,u_n\}$에 대해, LLM summarizer는 요약 항목(bullet) 집합 $B_t = \{b_1,...,b_m\}$을 생성하며, 이는 이후 메모리 관리 모델의 입력으로 사용된다. 요약 생성에 사용한 프롬프트는 <Figure 2>에 제시한다.

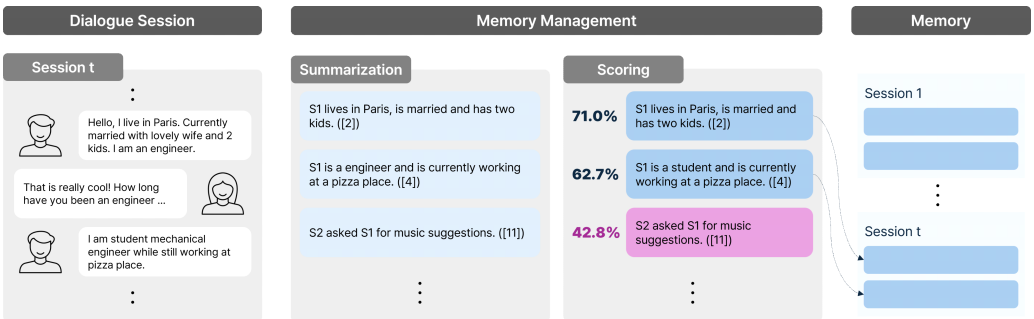

Figure 1. Overall Framework of the Proposed Methodology

```
You are summarizing a multi-session dialogue between Speaker 1 (S1) and Speaker 2 (S2).
Generate concise bullet points for THIS session only.

For EACH bullet, also return 1-3 evidence turns (indices) from the dialogue that directly support
the bullet.
Choose turns that contain the exact wording or facts reflected in the bullet.

Guidelines:
- Use only information explicitly stated in the session.
- Provide 3-8 bullets.
- Keep bullets brief and non-redundant.
- Each bullet should contain information that may be useful for maintaining continuity in FUTURE
sessions.
- Evidence indices must match the turn numbers in the dialogue.

Schema (follow exactly):
{
    "bullets": [
        {
            "text": "<bullet>", "evidence": [<turn_idx>, ...]
        }
    ]
}

Dialogue (turn-numbered):
{blocks}

Now generate the summary in the schema.
```

Figure 2. Prompt for Generating the Session Summary

3.2 메모리 관리

(1) 문제 정의

LLM 기반 요약은 종종 오류(hallucination)를 포함하거나, 중복되거나 비핵심적인 정보를 함께 생성하는 경향이 있다 (Belem *et al.*, 2025). 이러한 요약 항목들이 별도의 관리 없이 누적될 경우, 메모리 규모가 증가해 검색 및 응답 생성 단계에서 노이즈로 작용할 가능성이 크다. 실제로 일부 요약 항목은 후속 세션의 발화 생성에 실질적으로 기여하는 반면, 다른 항목은 문맥 의존적이거나 단발성 발화 수준에 머물러 이후 대화에서 재활용 가능성이 낮다. 반면 사용자의 활동이나 선호와 같이 이후 대화에서 재참조될 가능성이 높은 정보는 장기 대화 맥락에서 중요한 역할을 수행할 수 있다.

이러한 특성을 정성적으로 설명하기 위해, <Table 1>에는 저유용성 및 고유용성 요약 항목의 대표적인 예시를 제시한다. 의미가 불분명하거나 추상화가 충분히 이루어지지 않은 항목, 의미적 중요도가 낮은 단발성 발화 요약과 대비하여, 이후 세션에서 재참조 가능성이 높은 핵심 정보의 예시가 포함되어 있다. 이러한 사례는 요약 기반 메모리에서 모든 항목을 동일하게 유지하는 것이 비효율적일 수 있음을 보여준다.

이에 본 연구는 세션 요약이 생성된 이후, 각 요약 항목의 세션 간 유용성을 정량적으로 평가하고 이를 기반으로 항목을 선택적으로 유지하고 제거하는 메모리 관리 문제를 정의한다. 본 연구에서 유용성이란 특정 요약 항목이 다음 세션의 실제 발화 생성에 기여하는 정도를 의미한다. 메모리 관리 모델의 목표는 각 bullet b_i에 대해 이러한 유용성 점수 $y_i \in [0,1]$를 예측하는 scoring 함수 $f_\theta : b_i \mapsto \hat{y_i}$ 를 학습하는 것이며, 예측된 점수는 상위 k개(Top-k) 선택 또는 임계치(threshold) 기반 pruning에 활용된다.

또한 제안한 유용성 정의와 pruning 전략이 실제 대화 맥락에서 어떻게 작동하는지를 보이기 위해, 원 대화 일부와 해당 세션에서 생성된 요약 항목, 그리고 이에 대한 유용성 판단 유형을 함께 제시한 사례 분석을 부록 A에 포함하였다.

(2) LLM 기반 pseudo label 생성

학습 데이터 구축을 위해 본 연구에서는 LLM을 평가자로 활용하는 LLM-as-a-Judge 방식을 적용한다. 다음 세션 S_{t+1}의 실제 발화 a_{t+1}를 gold reference로 설정하고, 세션 S_t의 각 요약 bullet $b_i \in B_t$이 해당 발화를 생성하는 데 얼마나 기여하는지를 LLM에게 평가하도록 한다. LLM은 각 bullet에 대해 연속적인 유용성 점수 $y_{i,t} \in [0,1]$를 출력하며, 사용된 프롬프트는 <Figure 3>에 제시한다.

동일한 bullet은 다음 세션의 여러 턴에 대해 반복적으로 평가되므로, 하나의 bullet에 대해 복수의 점수 $\{y_{i,1}, y_{i,2}, \ldots\}$가 생성된다. 이를 세션 단위 bullet 라벨로 통합하기 위해 수식 (1)

Table 1. Representative Examples of Low Usefulness and High Usefulness Summary Bullets

Usefulness	Examples
Low	(*Ambiguous*) S2 is enthusiastic and would love to receive the link.
	(*Low semantic value*) S2 asked S1 whether they are a student.
	(Poor abstraction) S2 says “It’s not like I’m an alcoholic like my dad. I do drink from time to time but not to excess like him.”
High	S2 watched a movie at home with their partner and prefers horror films.
	S1 decided to join an intramural football league.

```
You are an evaluator that scores how useful each memory bullet is for generating the gold assistant
reply.

For each bullet, assign a real-valued importance score between 0.0 and 1.0:
- 0.0 = almost completely irrelevant
- 1.0 = crucial for generating the gold reply

Return ONLY valid JSON in the following format:
{{"scores": [s1, s2, ...]}}

The length of "scores" MUST be exactly equal to the number of bullets.
Do NOT include any extra keys, comments, or explanation.

Context:
{context}

Gold reply:
{gold}

Bullets (numbered):
{bullets}
```

Figure 3. Prompt for Judge-based Usefulness Scoring of Summary Bullets

과 같이 max pooling을 적용한다. 이는 평균 기반 집계에서 개별 턴의 핵심적 기여가 과도하게 희석되는 문제를 완화하고, 항목 단위의 최대 정보 기여를 보존하기 위한 집계 전략이다.

$$y_i = \max_t y_{i,t} \tag{1}$$

본 연구는 사용자 선호나 개인적 사실과 같이 장기 대화에서 정보의 중요도가 시간 지연을 두고 드러날 수 있다는 점을 고려한다. 이에 본 접근은 장기 지연 중요성을 명시적으로 모델링하기보다는, 세션 단위의 국소적 유용성(local usefulness) 추정이 반복적으로 누적될 경우 장기적 중요도에 대한 근사치를 제공할 수 있다는 운영적 가정을 따른다.

한편 pseudo label은 사람 어노테이터가 아닌 LLM에 의해 생성된다. 요약 항목의 기여도를 판단하는 문제는 본질적으로 반사실적(counterfactual) 추론을 요구하므로, 일관된 기준에 따라 대규모 인적 라벨을 구축하기 어렵다. 기존 연구에 따르면 적절한 프롬프트 설계 하에서 LLM은 의미적 관련성이나 선호 판단과 같은 평가 과제에서 신뢰 가능한 평가자로 기능할 수 있음이 보고되었다(Liu *et al.*, 2023; Rafailov *et al.*, 2023; Zheng *et al.*, 2023). 그러나 LLM의 판단 오류나 편향이 pseudo label 노이즈로 전이될 가능성은 존재하므로, 해당 점수를 정답 라벨이 아닌 약지도(weak supervision) 신호로 간주하고, 학습 과정 전반을 보수적으로 설계하였다. 이에 따른 구체적인 모델 구조와 단계적 학습 전략은 다음 절에서 상세히 설명한다.

3.3 메모리 관리 모델 구조

본 연구에서 제안하는 메모리 관리 모델은 LLM-as-a-Judge 방식으로 생성된 pseudo label을 약지도 신호로 활용하는 학습 환경을 전제로 설계되었다. 특히 pseudo label에 내재할 수 있는 노이즈와 불확실성을 고려하여, 사전학습 언어모델이 제공하는 고품질 의미 표현을 최대한 보존하면서도 항목 단위의 유용성 판단 기능을 안정적으로 학습할 수 있도록 모델 구조를 구성하였다.

이를 위해 학습은 두 단계로 수행된다. 1단계에서는 인코더를 완전히 동결한(frozen) 상태에서 경량 예측 모듈만 학습하여 기본적인 유용성 판단 기능을 안정적으로 습득한다. 2단계에서는 LoRA 기반의 제한적 인코더 튜닝을 적용하여, 사전 학습된 의미 표현을 크게 훼손하지 않으면서 과제에 필요한 의미적 구분을 미세하게 반영한다. 이러한 단계적 학습 전략은 pseudo label 기반 약지도 학습 환경에서 라벨 노이즈로 인한 과적합 위험을 줄이고, 노이즈가 인코더의 의미 표현 공간 전체로 확산되는 것을 방지하기 위한 설계이다.

(1) 인코더 기반 경량 예측 구조

사전학습 언어 모델 기반의 문장 임베딩 모델을 요약 항목 인코딩에 활용한다. 이러한 모델은 대규모 코퍼스 학습을 통해 문장 간 의미적 유사성(semantic similarity)을 효과적으로 포착하며(Reimers *et al.*, 2019), 요약 항목과 같은 짧은 텍스트 단위에서도 안정적인 의미 표현을 제공한다. 이에 따라 1단계의 목표는 인코더 자체를 재학습하는 것이 아니라, 사전학습 임베딩 위에서 유용성 점수를 예측하는 얕은 mapping 함수를 학습하는 것이다.

이 단계에서 인코더는 학습 전반에 걸쳐 완전히 동결되며, 인코더 출력 위의 projection layer와 scoring layer만을 업데이트한다. Bullet b_i의 인코딩은 수식 (2)와 같이 계산된다.

$$h_i = \mathrm{Encoder}(b_i) \tag{2}$$

이후 projection 및 scoring 과정을 거쳐 예측 유용성 점수 $\hat{y}_i$가 산출된다.

$$z_i = W_p h_i + b_p, \quad \hat{y}_i = \sigma(W_s z_i + b_s) \tag{3}$$

실제 구현에서는 projection layer 뒤에 Layer Normalization을 적용하여 학습 안정성을 높였으며, scoring layer는 bias 없는 단일 차원 선형 변환으로 구성하였다. 여기서 σ는 sigmoid 함수이며, $\hat{y}_i$는 bullet b_i에 대한 예측 유용성 점수를 의미한다.

$$z_i = \mathrm{LayerNor}(\mathrm{W}h_i) \tag{4}$$

$$\hat{y}_i = \sigma(w^\top z_i) \tag{5}$$

동결 인코더 구조는 본 연구의 데이터 조건 및 학습 환경과 잘 부합한다. 인코더 파라미터를 고정함으로써 학습해야 할 파라미터 수가 크게 감소하여 학습이 빠르고 안정적으로 진행되며, pseudo label 기반의 제한된 데이터 환경에서 인코더 전체를 업데이트할 때 발생할 수 있는 과적합 위험을 완화할 수 있다. 이러한 설계는 사전학습 모델이 포착한 고품질의 일반적인 의미 표현은 유지하면서, 유용성 예측에 필요한 최소한의 mapping 기능만을 학습하도록 한다.

(2) LoRA 기반 제한적 인코더 튜닝

1단계의 Head-only 학습은 계산 효율성과 학습 안정성 측면에서 유리하지만, 항목 단위 유용성 판단처럼 미세한 의미 차이를 요구하는 과제에서는 표현력이 제한될 수 있다. 이를 보완하기 위해 2단계에서는 LoRA(Low-Rank Adaptation)(Hu *et al.*, 2022)를 적용하여 제한적으로 인코더 튜닝을 수행한다.

LoRA는 인코더의 기본 가중치(base weight)를 유지한 채, 셀프 어텐션(self-attention) 모듈의 query, key, value 경로에 저랭크(low-rank) 행렬을 삽입하고 해당 파라미터만 학습하는 방식이다. 이를 통해 전체 파라미터를 업데이트하지 않고도 과제 특화 표현을 보완할 수 있으며, 사전학습 표현을 과도하게 훼손하지 않는 장점이 있다. 본 연구에서는 안정성을 강화하기

위해 2단계에서 1단계보다 더 작은 학습률을 사용하였다. 결과적으로 제안한 2단계 학습은 의미 표현의 일반성을 유지하면서도 항목 단위 판별에 필요한 과제 특화 조정을 제한적으로 반영하도록 설계되었다.

3.4 학습 손실 함수

본 연구는 회귀적 정확도와 랭킹 정합성을 동시에 확보하기 위해 회귀 손실(regression loss)과 쌍별 랭킹 손실(pairwise ranking loss)을 결합한 다목적 학습(multi-objective training)을 적용하였다.

(1) 회귀 손실(Regression Loss)

예측된 유용성 점수가 실제 라벨과 근접하도록 하기 위해 수식 (6)과 같이 smooth L1 loss(Girshick, 2015)를 사용한다.

$$L_{reg} = \mathrm{SmoothL1}(\hat{y_i}, y_i) \tag{6}$$

이러한 회귀 손실은 모델이 예측한 유용성 점수 $\hat{y_i}$를 실제 라벨 y_i의 스케일에 맞게 정렬시키는 역할을 수행하며, 학습 과정의 안정적 수렴을 유도한다. 특히 smooth L1 loss는 이상치(outlier)에 덜 민감한 강건한 회귀 손실로, 약지도 환경에서 발생할 수 있는 라벨 노이즈의 영향을 완화하고 학습 초기의 불안정성을 줄이는 데 유리하다(Terven *et al.*, 2025). 한편, 쌍별 랭킹 손실은 항목 간 상대적 순위를 학습하는 데 효과적이지만 점수의 절대적인 보정(calibration)을 보장하지 못한다(Cao *et al.*, 2007; Menon *et al.*, 2012). 이러한 이유로 회귀와 랭킹 목표를 동시에 최적화하는 결합 학습이 순위 성능과 점수 안정성을 모두 향상시킨다는 연구가 보고되어 왔다(Sculley, 2010; Yan *et al.*, 2022; Bai *et al.*, 2023). 이에 본 연구에서도 두 손실을 함께 사용함으로써 절대적 scoring 능력(regression)과 상대적 중요도 판단(ranking)을 동시에 학습하도록 설계하였다.

(2) 쌍별 랭킹 손실(Pairwise Ranking Loss)

동일 세션 내 $y_i > y_j$인 bullet 쌍에 대해, 다음의 수식 (7)과 같은 margin-ranking loss를 적용한다. 이는 모델이 bullet 간 상대적 중요도 순위 정보(ranking)를 학습하도록 돕는다.

$$L_{rank} = \max(0, m - (\hat{y_i} - \hat{y_j})), \quad m = 0.05 \tag{7}$$

(3) 최종 손실 함수

수식 (8)과 같이 두 손실의 가중 합으로 최종 손실 함수를 구성한다. 회귀 손실은 점수 보정을, 랭킹 손실은 상대적 순위 학습을 담당하여, 두 목표를 동시에 최적화함으로써 bullet 단위 pruning 성능을 개선한다.

$$L = L_{reg} + \lambda L_{rank} \tag{8}$$

실험에서는 $\lambda = 0.7$로 설정하여 회귀 안정성과 순위 정합성의 균형을 맞추었다.

4. 실 험

4.1 실험 설정

문장 임베딩 모델로 널리 활용되는 여러 사전학습 모델을 대상으로 비교 실험을 수행하였다. 구체적으로는 E5-Large (*intfloat/e5-large-v2*)과 E5 Multilingual(*intfloat/multilingual-e5-large*)(Wang *et al.*, 2024), Qwen3-0.6B 임베딩(*Qwen/Qwen3-Embedding-0.6B*)(Zhang *et al.*, 2025), 그리고 BGE-Large-En (*BAAI/bge-large-en-v1.5*)(Xiao *et al.*, 2024)을 후보로 선정하였다. 이들 후보 모델 중 유용성 예측 성능이 가장 우수하게 나타난 E5 Multilingual 모델을 최종 실험에 사용하였으며, 후보별 성능 비교 결과는 부록 B의 <Table 8>에 제시하였다. 모델 학습은 두 단계 모두 5 epoch로 수행하였고, batch size는 32로 설정하였다. 학습률(learning rate)은 1단계에서 1e-4, 2단계에서 5e-5를 사용하였다.

실험에서 활용한 LLM은 단계별로 상이하다. 모두 GPT 계열 모델(Achiam *et al.*, 2023)을 활용하였는데, 세션 요약 생성 단계에서는 GPT-5-mini(*gpt-5-mini*) API를 사용하였으며, 최종 대화 응답 생성 단계에서는 GPT-4.1(*gpt-4.1*)을 적용하였다.

4.2 데이터셋

모델 학습 및 평가는 MSC(Multi-Session Chat) 데이터셋(Xu *et al.*, 2022)을 활용하였다. MSC는 PersonaChat(Zhang *et al.*, 2018) 기반의 페르소나를 가진 두 화자(Speaker 1, Speaker 2)가 몇 시간에서 며칠에 걸친 시간 간격을 두고 여러 세션에 걸쳐 수행한 장기 대화로 구성된다. 각 대화는 최대 5개 세션으로 이루어지며, 세션당 최대 14개 발화를 포함한다.

Pseudo label 생성 과정에서의 LLM 호출 비용을 고려하여, MSC 데이터셋의 valid set에 포함된 500개 전체 대화를 모델 학습에 사용하였다. 모델 성능 평가는 MSC 데이터셋의 test set 중 100개 대화를 대상으로 수행하였다.

4.3 평가 방법

(1) 모델 성능 평가

학습을 수행하지 않은 Vanilla 모델, head-only 미세조정을 적용한 1단계(Stage 1) 모델, 그리고 LoRA 기반 제한적 인코더 튜닝을 적용한 2단계(Stage 2) 모델의 세 가지 설정을 비교하였다. 각 모델은 동일한 pseudo label 기반 평가 데이터셋에서

항목별 유용성 점수를 예측하고, 이를 pseudo label과 비교하여 성능을 분석하였다. 회귀 기반 지표와 랭킹 기반 지표를 활용하여 정량적으로 평가하였으며, 지표의 정의는 4.4절에 제시한다. 이를 통해 각 학습 단계가 메모리 유용성 예측 성능에 미치는 영향을 체계적으로 검증하였다.

(2) 메모리 효율성 평가

메모리 관리 모델 적용으로 대화 모델에 전달되는 메모리 입력이 얼마나 감소하는지 정량적으로 평가하였다. Pruning 적용 전후를 비교하여 (i) 세션당 유지되는 메모리 항목 수와 (ii) 대화 모델 입력에 포함되는 전체 토큰 수를 측정하였다. 입력 토큰 수는 LLM 기반 대화 시스템에서 추론 비용 및 응답 지연(latency)과 밀접하게 연관되므로, 시스템 관점의 효율성을 반영하는 핵심 지표로 간주하였다. 또한 상위 후보 개수(top-k)에 따른 효율성 변화를 함께 분석하여 메모리 효율성과 성능 간의 trade-off를 평가하였다.

(3) 대화 성능 평가

메모리 pruning이 대화 응답 생성 품질에 미치는 영향을 평가하기 위해, 서로 다른 메모리 제공 방식에 따른 세 가지 조건을 설정하였다. 첫 번째는 대화 모델에 어떠한 메모리도 제공하지 않는 조건(No memory)이며, 두 번째는 pruning을 적용하지 않은 전체 요약 메모리를 그대로 제공하는 조건(All memory), 세 번째는 학습된 메모리 관리 모델이 선별한 항목만을 제공하는 조건(Pruned memory)이다. Pruned memory 조건에서는 메모리 관리 모델이 예측한 유용성 점수에 대해 threshold로 1차 필터링한 뒤, 남은 후보 중 상위 k개 항목(top-k)만 유지하였다. 기본 설정은 threshold = 0.5, k = 5이다.

동일한 입력에 대해 생성된 응답을 원 데이터셋의 gold reference와 비교하였으며, test 데이터셋의 마지막 세션(세션 5)에서 무작위로 선택한 50개 샘플을 대상으로 평가하였다. 비교 대상 모델로는 LLM을 활용하여 세션 요약 또는 메모리 구성을 수행하는 기존 방법론 중, 세션 5 결과를 명시적으로 보고하고 있어 공정한 비교가 가능한 RecurSum(Wang *et al.*, 2025)과 LD-Agent(Li *et al.*, 2025)를 포함하였다. 또한 THEANINE(Ong *et al.*, 2025)은 세션별 성능이 아닌 세션 전반에 대한 평균 성능만을 보고하므로, 본 연구에서는 THEANINE의 보고된 평균 성능을 참고 지표로만 제시하며 세션 5에 대한 직접 비교로 해석하지 않는다.

4.4 평가 지표

(1) 모델 성능 평가

모델 평가에는 회귀 성능 지표와 랭킹 기반 지표를 함께 사용하여, 메모리 관리 모델의 정량적 성능을 다각도로 평가하였다.

회귀 성능 평가 지표: MSE(Mean Squared Error), MAE(Mean Absolute Error), Pearson 상관계수, Spearman 순위 상관계수를 사용하였다. MSE는 제곱 오차의 평균, MAE는 절대 오차의 평균으로 예측 오차 크기를 측정한다. Pearson 상관계수는 예측값과 실제값 간의 선형 상관관계를 측정하고(Benesty *et al.*, 2009), Spearman 순위 상관계수는 두 변수 간 순위 관계의 일치도를 평가한다(Spearman, 1904). 본 연구에서는 Spearman 계수를 전역 및 세션 단위에서 모두 산출하였다.

세션 단위 랭킹 성능 평가 지표: 세션 내 여러 메모리 후보 항목 중에서 모델이 중요한 항목을 얼마나 정확하게 식별하는지를 측정하기 위해 NDCG@k, Hit@k, Precision@k, Recall@k 지표를 사용하였다. NDCG@k는 상위 k개 예측 결과가 실제 중요도를 얼마나 잘 반영하는지 평가하며(Jarvelin *et al.*, 2002), Hit@k는 세션 내 가장 중요한 항목이 상위 k개 예측 결과에 포함되는지를 측정한다(He *et al.*, 2017). Precision@k와 Recall@k는 실제로 중요한 메모리 항목을 모델이 상위 k개 예측 결과에서 얼마나 정확하게 회수하는지를 정량화하는 지표이다(Koren *et al.*, 2009).

본 연구에서는 Precision@k 및 Recall@k 계산을 위해 중요 항목을 정의하는 라벨 임계값을 0.6으로 설정하였다. 해당 임계값은 연속적인 중요도 라벨을 이진화하기 위한 평가 기준으로, pruning 과정에서 사용되는 예측 점수 필터(threshold=0.5)와는 목적이 다르다. 이는 중요 항목을 보수적으로 정의하여 평가 안정성을 확보하기 위한 설정으로, 0.5 이상의 임계값 구간에서 주요 랭킹 지표의 상대적 경향이 유지됨을 사전에 확인하였다.

통계 검정: 모델 간 성능 차이의 유의성을 검증하기 위해 동일 데이터 분할에서 서로 다른 세 개의 랜덤 시드로 실험을 반복하고 평균과 표준편차(mean ± std)로 보고하였다. 주요 비교는 Vanilla와 최종 모델(Stage 2)로 설정하였다. 전역 회귀 지표는 대응 표본 양측 t-검정(paired two-sided t-test)을 적용하였고, 세션 단위 랭킹 지표는 비정규 가능성을 고려하여 Wilcoxon signed-rank test를 사용하였다. 유의수준은 0.05로 설정했으며, 지표별 p-value는 부록 C의 <Table 9>에 제시하였다.

(2) 대화 성능 평가

대화 응답의 정량 평가는 BLEU, ROUGE-L, BERTScore, MAUVE의 자동 평가 지표를 사용하였다. BLEU(Papineni *et al.*, 2002)는 n-gram 정밀도에 기반하여 생성 문장이 참조 문장에서 나타나는 n-gram을 얼마나 재현하는지를 측정하는 지표이다. ROUGE-L(Lin, 2004)은 생성 문장과 참조 문장 간의 최장 공통 부분수열을 기반으로 유사도를 평가하는 지표이다. BERTScore(Zhang *et al.*, 2020)는 사전학습 언어모델이 생성하는 문맥 임베딩을 활용하여 생성 문장과 참조 문장 간의 의미적 정밀도와 재현율을 측정하는 지표이다. 마지막으로, MAUVE(Pillutla *et al.*, 2021)는 생성 분포와 참조 분포 간의 전반적 유사도를 f-divergence 기반으로 측정하는 지표로, 최근 대규모 언어 모델의 응답 품질 평가에서 활발하게 활용되고

있다. MAUVE는 응답 품질뿐 아니라 다양성과 자연스러움까지 분포 수준에서 포착할 수 있어, 대화 모델 평가에 특히 적합한 특성을 갖는다.

5. 실험 결과 및 분석

5.1 모델 성능 평가

본 절에서는 제안한 메모리 관리 모델의 성능을 다중 랜덤 시드(multi-seed) 환경에서 평가하고, 회귀 지표와 세션 단위 랭킹 지표에 대해 통계적 유의성 검정 결과를 분석한다. 전역 회귀 성능은 <Table 2>에, 세션 단위 랭킹 성능은 <Table 3>에 각각 평균 ± 표준편차(mean ± std) 형태로 제시하였다. 또한, 상위 후보 개수 k 변화에 따른 Stage 2 모델의 랭킹 성능 특성은 추가 분석으로 제시한다.

(1) 회귀 기반 성능 분석

<Table 2>는 Vanilla 모델, Stage 1, Stage 2 모델의 전역 회귀 성능을 비교한 결과를 나타낸다. Stage 2 모델은 모든 회귀 지표에서 가장 낮은 예측 오차를 기록하였으며, 이는 Vanilla 모델 대비 각각 약 8.3%, 7.9%의 상대적 감소에 해당한다. Stage 1 모델 역시 Vanilla 대비 유의미한 오차 감소를 보였으나, Stage 2 모델은 이를 추가적으로 개선하였다. 이는 LoRA 기반 인코더 튜닝이 항목 단위 유용성 예측의 정밀도를 보완하는 데 효과적임을 시사한다.

Table 2. Global regression performance of memory scoring models (mean ± std over three seeds). * denotes statistically significant improvement over the Vanilla model ($p < 0.05$)

Metric	Vanilla Model	Fine-tuned Model	
		Stage 1	Stage 2
MSE	0.1915 (±0.0013)	0.1773 (± 0.0004)	0.1755 (± 0.0022)*
MAE	0.4168 (±0.0012)	0.3926 (± 0.0005)	0.3837 (± 0.0017)*
Pearson Correlation	0.0038 (±0.0124)	0.2449 (± 0.0027)	0.2589 (± 0.0026)*
Spearman Correlation	0.0101 (±0.0153)	0.2522 (± 0.0020)	0.2686 (± 0.0020)*

상관 지표에서도 일관된 개선이 관찰되었다. Pearson 및 Spearman 상관계수는 Vanilla 모델 대비 Stage 1과 Stage 2로 갈수록 단계적으로 증가하였으며, 이는 모델이 유용성 라벨의 전역적 분포 구조와 항목 간 상대적 순위 관계를 점진적으로 더 정확하게 학습하고 있음을 보여준다.

통계적 유의성 검증 결과, Stage 2 모델은 Vanilla 모델 대비 모든 회귀 지표에서 유의미한 성능 향상을 보였다($p < 0.05$). 특히 Pearson 및 Spearman 상관계수는 $p < 0.01$ 수준의 유의성을 보여, 예측 오차 감소뿐만 아니라 유용성 점수의 전반적인 순위 구조를 보다 안정적으로 학습하고 있음을 확인할 수 있다. 또한 인코더를 완전히 동결한 Stage 1 모델 역시 Vanilla 대비 일관된 개선을 보였다는 점은, pseudo label 기반 약지도 환경에서도 보수적인 head-only 학습 전략이 효과적으로 작동함을 시사한다.

(2) 세션 단위 랭킹 성능 평가 분석

<Table 3>은 k = 5 기준에서 세션 단위 랭킹 성능을 비교한 결과를 제시한다. 세션 단위 Spearman 상관계수는 Vanilla 모델의 -0.0011에서 Stage 1의 0.2422, Stage 2의 0.2790으로 크게 향상되었으며, 이는 Stage 2 모델이 동일 세션 내 메모리 후보들 간 상대적 중요도 순위를 가장 정확하게 정렬하고 있음을 의미한다. 해당 개선은 Wilcoxon signed-rank test를 통해 통계적으로 유의함이 확인되었다($p < 0.001$).

Table 3. Session-level ranking performance of memory scoring models (mean across sessions). * denotes statistically significant improvement over the Vanilla model ($p < 0.05$)

Metric	Vanilla Model	Fine-tuned Model	
		Stage 1	Stage 2
Spearman	-0.0011 (±0.0271)	0.2422 (±0.0032)	0.2790 (±0.0053)*
NDCG@5	0.5909 (±0.0190)	0.7196 (±0.0003)	0.7297 (±0.0009)*
Hit@5	0.7100 (±0.0337)	0.8567 (±0.0042)	0.8733 (±0.0068)*
Precision@5	0.3746 (±0.0122)	0.4377 (±0.0003)	0.4403 (±0.0015)*
Recall@5	0.7059 (±0.0297)	0.8482 (±0.0010)	0.8510 (±0.0045)*

NDCG@5 또한 Vanilla 대비 Stage 1과 Stage 2로 갈수록 점진적으로 증가하였으며, Stage 2의 개선은 통계적으로 유의하였다($p < 0.01$). 이는 LoRA 기반 인코더 튜닝을 통해 모델이 실제로 중요한 메모리 항목을 상위 순위에 보다 안정적으로 배치할 수 있게 되었음을 시사한다. Hit@5 역시 Stage 2에서 크게 향상되어, 세션 내 가장 중요한 항목을 높은 확률로 상위 5개 후보 내에서 복원할 수 있음을 보여준다($p < 0.001$). Precision@5와 Recall@5 또한 Stage 2 모델이 가장 우수한 성능을 기록하였으며, 이는 pruning 과정에서 핵심 정보의 누락을 줄이면서 불필요한 메모리를 효과적으로 제거할 수 있음을 의미한다.

본 과제는 pseudo label 기반 약지도 환경에서 미세한 상대적

중요도 판단을 요구하므로, 수치적으로는 제한적인 차이라 하더라도 통계적으로 일관된 성능 향상은 실제 메모리 pruning 결정의 신뢰도를 실질적으로 개선한다. 랭킹 지표 전반에서 관찰된 이러한 일관된 개선은, LLM-as-a-Judge가 제공한 감독 신호가 상대적 중요도 학습 측면에서 충분한 일관성을 가지며, 실제 pruning 시나리오에서도 유효하게 활용될 수 있음을 시사한다.

(3) k 값 변화에 따른 Stage 2 랭킹 성능 분석

Stage 2 모델의 상위 후보 개수 k 변화에 따른 랭킹 성능을 분석하기 위해 k = 1, 3, 5 조건에서 주요 지표를 비교한 결과를 <Figure 4>에 제시하였다. 각 k 값에 대한 세부 수치와 모든 지표의 측정 결과는 부록 D의 <Table 10>에 제시하였다. 분석 결과, NDCG@k와 Hit@k는 k 증가에 따라 일관되게 향상되어, 유지되는 후보 수가 증가할수록 실제로 중요한 메모리 항목이 상위 집합에 포함될 가능성이 높아짐을 확인하였다.

반면 Precision@k는 k 증가에 따라 점진적으로 감소하는 경향을 보였으나, Recall@k는 크게 상승하였다. 이는 소수의 핵심 항목만을 유지하는 설정에서는 정밀도가 높지만 정보 누락 위험이 크고, 반대로 k가 증가할수록 대부분의 유용한 항목이 안정적으로 회수됨을 보여준다. 종합하면 Stage 2 모델은 k 값 변화에 따라 랭킹 지표 전반에서 일관된 성능 향상을 유지하였으며, 특히 k = 5 설정은 정밀도와 재현율 간의 균형이 가장 안정적으로 유지되는 구간으로 성능 안정성과 메모리 예산을 균형 있게 만족하는 실용적인 기본 설정임을 시사한다.

(4) 종합 분석

전역 회귀 지표와 세션 단위 랭킹 지표를 종합한 결과, Stage 1 모델은 head-only 튜닝만으로도 pseudo label 기반 약지도 환경에서 Vanilla 모델 대비 안정적인 성능 향상을 달성하였다. Stage 2 모델은 LoRA 기반 제한적 인코더 조정을 통해 이러한 성능을 추가적으로 정밀화하였으며, 전역 오차 감소, 상관 지표 향상, 그리고 세션 단위 핵심 항목 복원 능력의 세 측면에서 모두 통계적으로 유의미한 개선을 보였다.

일부 지표에서 절대적인 성능 향상 폭은 제한적으로 보일 수 있으나, 관측된 개선은 다양한 평가 지표와 다중 랜덤 시드 실험 전반에서 일관되게 재현되었고, 통계 검정을 통해 그 유의성이 확인되었다. 이는 제안한 단계적 학습 전략이 단순한 성능 변동이 아니라, 유용성 점수 예측과 항목 간 상대적 중요도 판단 능력을 구조적으로 개선함을 시사한다.

이러한 결과는 본 연구의 접근법이 LLM 기반 pseudo label에 의존하는 약지도 학습 환경이라는 제약 하에서도, 실제 메모리 관리 및 pruning 응용 시나리오에서 실용적인 성능을 제공할 수 있음을 보여준다. 해당 특성은 이후 절의 대화 성능 평가 결과에서도 일관되게 확인된다.

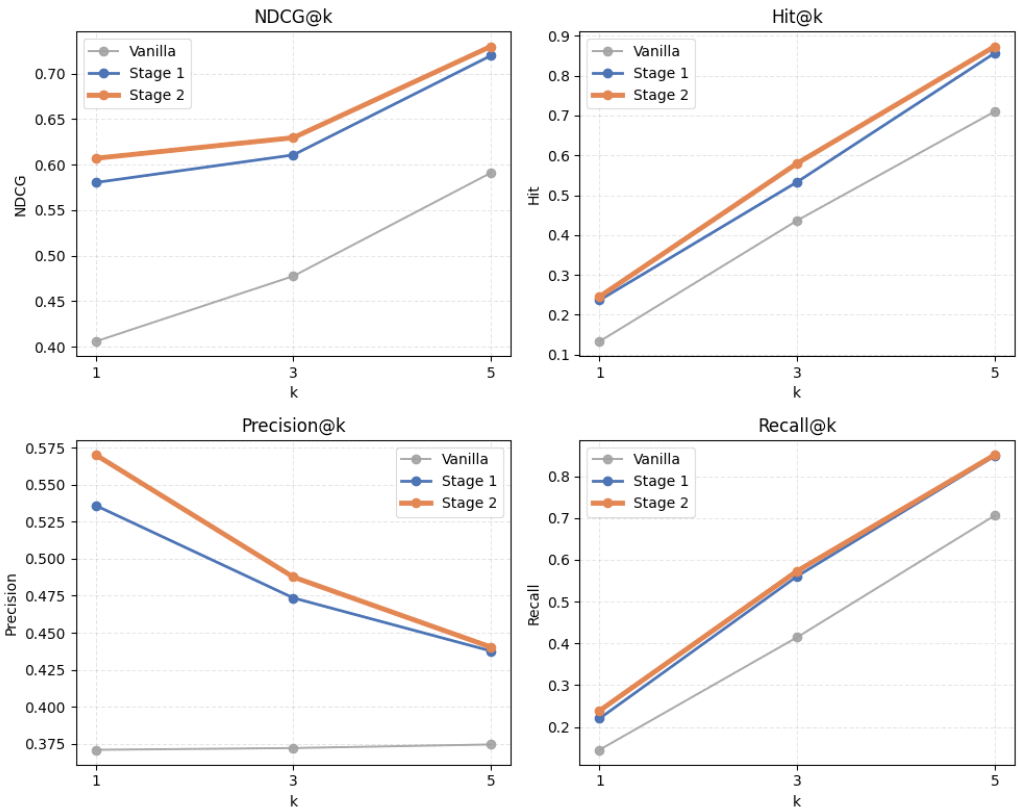

Figure 4. Session-level Ranking Performance of the Stage 2 Model Under Different k Values

5.2 메모리 효율성 분석

본 절에서는 제안한 메모리 관리 모델이 대화 시스템 환경에서 메모리 사용량과 입력 토큰 수를 얼마나 감소시키는지를 정량적으로 분석한다. 예측 유용성 점수에 대한 threshold는 신뢰도가 낮은 항목을 사전에 제거하기 위한 품질 필터로 사용되며, 상위 항목 개수(top-k)는 필터링 이후 최종적으로 유지되는 항목 수를 제한하는 예산(budget) 제어 변수로 작동한다. 따라서 두 변수는 대체 관계가 아니라 pruning 과정에서 상호 보완적으로 작동한다.

Threshold 설정의 타당성을 검증하기 위해 부록 E에서 threshold 변화에 따른 민감도 분석을 제시한다. 분석 결과, threshold가 증가할수록 유지되는 항목 수가 급격히 감소하여 pruning 강도가 크게 강화되는 경향이 관찰되었다. 이에 따라 메모리 효율성과 정보 보존 간의 균형이 비교적 안정적으로 유지되는 threshold = 0.5를 기본 설정으로 선택하고, 이후 본 절의 모든 효율성 분석에서는 해당 값을 고정하였다.

한편 top-k는 유지되는 항목 수를 직접 제어하므로, 효율성과 성능 간 trade-off에 보다 직접적인 영향을 미친다. 이에 따라 5.1절의 k 변화에 따른 랭킹 성능 분석과 본 절의 효율성 분석을 함께 고려하여 top-k 설정을 종합적으로 논의한다.

(1) 항목 수준 메모리 감소 분석

<Table 4>는 threshold를 0.5로 고정한 상태에서 top-k 설정에 따른 세션당 평균 유지 메모리 항목 수와 감소율을 나타낸다. 제안한 방법은 보수적인 설정(top-k = 5) 하에서도 세션당 유지 메모리 항목 수를 평균 6.95개에서 2.69개로 줄여 약 61%의 항목 감소를 달성하였다. Top-k = 3 설정은 평균 2.31개로 더 높은 압축률을 보였으나, 앞선 성능 분석에서 확인된 바와 같이 중요 항목 복원 성능의 안정성이 함께 저하될 가능성이 존재한다. Top-k = 1 설정은 가장 높은 감소율을 제공하지만, 단일 항목만을 유지하는 경우 장기 대화 맥락에서 요구되는 정보 다양성을 충분히 보존하기 어렵다.

Table 4. Bullet-level memory reduction under different top-k settings (threshold=0.5)

Setting	Avg. # Bullets	Bullet Reduction (%)
No pruning	6.95	
Proposed Method (Top-K=5)	2.69	61.36
Proposed Method (Top-K=3)	2.31	66.77
Proposed Method (Top-K=1)	0.96	86.21

(2) 토큰 수준 메모리 감소 분석

<Table 5>는 동일한 threshold 설정(threshold = 0.5) 하에서 top-k 변화에 따른 입력 토큰 수 및 감소율을 나타낸다. Top-k = 5 설정에서도 전체 입력 토큰 수가 약 60% 감소하여, 제안한 메모리 관리 방법이 실제 대화 시스템의 입력 비용 및 추론 부담을 실질적으로 완화할 수 있음을 확인하였다. Top-k = 3 설정에서는 토큰 감소율이 더 높게 나타났으나, 이는 중요한 정보가 함께 제거될 가능성이 증가하는 설정으로 해석할 수 있다. 한편 top-k = 1 설정은 가장 높은 토큰 감소율을 제공하지만, 앞선 성능 분석에서 확인된 바와 같이 핵심 정보 보존 측면에서는 한계를 보인다.

Table 5. Token-level Memory Reduction Under Different top-k Settings (threshold=0.5)

Setting	Total # Tokens	Token Reduction (%)
No pruning	265,062	
Proposed Method (Top-K=5)	107,129	59.58
Proposed Method (Top-K=3)	92,081	65.26
Proposed Method (Top-K=1)	40,689	84.65

(3) 종합 분석

항목 수준 및 토큰 수준 분석을 종합하면, 작은 top-k 값은 높은 압축 효율을 제공하는 반면 정보 손실 위험을 증가시키는 경향이 있다. 반대로 top-k = 5 설정은 약 60% 수준의 메모리 및 토큰 절감을 달성하면서도, 핵심 정보 복원 성능이 가장 안정적으로 유지되는 구간에 해당한다. 이에 따라 본 연구에서는 효율성과 성능 안정성 간의 균형을 고려하여 threshold = 0.5, top-k = 5를 기본 설정으로 채택하였다.

5.3 대화 성능 평가

<Table 6>은 서로 다른 메모리 조건에서 생성된 응답을 자동 평가 지표로 비교한 결과를 보여준다. 전반적으로 Pruned memory 조건이 주요 지표 전반에서 가장 높은 성능을 기록하여, 메모리 pruning이 응답 품질 향상에 기여함을 확인할 수 있다. Pruned memory 조건에서의 개선은 중복되거나 불필요한 정보가 제거되면서 모델이 핵심 정보에 더 집중할 수 있었기 때문으로 해석된다. 특히 All memory 조건의 성능이 Pruned memory보다 낮게 나타난 결과는, 과도한 메모리 제공이 오히려 정보적 잡음을 유입하여 응답 생성에 부정적인 영향을 줄 수 있음을 시사한다.

또한 LD-Agent와 같은 메모리 검색(retrieval) 기반 접근과 비교했을 때, 본 연구의 pruning 기반 접근은 별도의 검색 단계 없이도 경쟁력 있는 성능을 보였다. 자동 평가 지표가 검색 기반 접근의 장점을 완전히 포착하지 못할 가능성을 고려하더라도, 대화 생성 성능에서 메모리의 양보다 제공되는 정보의 품질이 중요한 요인으로 작용할 수 있음을 뒷받침한다.

이러한 결과는 요약 메모리를 단순히 확장하기보다 항목 단위로 품질을 관리하여 핵심적이고 관련성 높은 정보만 유지하는 전략이 대화 응답 품질 향상에 효과적임을 확인하였다. 특

Table 6. Automatic evaluation results (%) across different memory conditions

Metric	B-1	B-2	B-3	B-4	R-L	BERTScore	MAUVE
Baselines							
RecurSum	**21.83**	**12.59**			17.86	**86.89**	
LD-Agent		7.37	3.03		15.17		
THEANINE				1.8	15.37	86.70	18.62
Our Approach							
No memory	18.65	6.69	3.19	1.70	15.52	85.87	15.91
All memory	20.06	7.86	3.68	1.77	16.94	86.43	16.70
Pruned memory	20.70	8.63	**4.23**	**2.04**	**17.59**	86.54	**34.97**

히 MAUVE 점수의 큰 향상은 pruning이 불필요하거나 잡음에 해당하는 메모리 정보를 억제하는 동시에, 누적된 메모리를 무차별적으로 제공할 때 발생하기 쉬운 응답 분포의 왜곡을 완화하여 보다 균형 있고 다양한 응답 분포를 유지하는 데 기여함을 시사한다.

6. 결 론

본 연구는 LLM 기반 장기 대화 시스템에서 세션 요약에 포함된 개별 항목의 유용성을 정량적으로 평가하고, 중요도가 낮은 정보를 효과적으로 제거하기 위한 요약 기반 메모리 관리 프레임워크를 제안하였다. 이를 위해 LLM-as-a-Judge 방식을 활용하여 pseudo label을 생성하고, 사전학습 인코더를 기반으로 한 head-only 학습 단계와 LoRA를 적용한 제한적 인코더 미세 조정 단계로 구성된 2단계 학습 전략을 설계하였다. 이러한 접근은 경량 구조를 유지하면서도 항목 단위 유용성 예측에 필요한 표현을 효과적으로 학습할 수 있게 하였으며, 기존 연구에서 상대적으로 충분히 다뤄지지 않았던 요약 항목 단위의 메모리 관리 문제를 체계적으로 접근할 수 있는 방법을 제시하였다.

회귀 및 세션 단위 랭킹 평가 결과, 제안한 모델은 항목 단위 유용성 분포와 상대적 중요도 구조를 보다 정밀하게 포착하였으며, pruning 환경에서도 핵심 정보를 안정적으로 보존하는 성능을 보였다. 또한 메모리 및 입력 토큰 수를 크게 감소시키면서도 대화 응답 품질을 개선함으로써, 메모리의 양보다 제공되는 정보의 질이 대화 성능에 중요한 영향을 미친다는 점을 확인하였다.

대화 생성 평가에서도 pruning된 메모리를 제공한 조건이 BLEU, ROUGE-L, BERTScore, MAUVE 등 다양한 자동 평가 지표에서 일관되게 우수한 성능을 기록하였다. 이는 불필요하거나 중복된 정보를 포함한 메모리를 그대로 활용하는 것보다, 요약 항목의 품질을 사전에 관리하여 핵심적이고 관련성 높은 정보만을 유지하는 접근이 LLM의 응답 생성 품질을 효과적으로 향상시킨다는 것을 의미한다. 특히 검색 모듈을 포함한 기존 기법과 비교했을 때, 본 연구의 pruning 기반 접근이 검색 절차 없이도 경쟁력 있는 성능을 보였다는 점은 장기 대화 시스템에서 메모리 관리의 핵심이 정보 선택과 정제에 있음을 시사한다.

한편 본 연구에는 몇 가지 한계가 존재한다. 첫째, 요약 항목의 유용성을 다음 세션 기여도 기준으로 정의하여 장기 지연 중요성을 명시적으로 모델링하지는 못하였다. 둘째, pseudo label은 LLM의 판단에 의존하므로 라벨 품질이 모델 성능의 상한을 제한할 수 있다. 셋째, LLM 호출 비용 제약으로 학습 데이터 규모가 제한되어, 데이터 확장에 따른 성능 추세에 대한 추가 검증이 필요하다. 넷째, 메모리 검색을 사용하지 않는 설정을 가정하였기 때문에 pruning과 검색 모듈이 결합된 환경에서의 상호작용을 충분히 고려하지 못했다. 마지막으로 자동 평가 지표를 중심으로 분석을 수행하여 실제 사용자 경험을 반영한 정성적 평가가 부족하다.

향후 연구에서는 소량의 고품질 인적 라벨과 pseudo label을 결합한 반지도 학습이나, 다중 LLM 평가자 간 합의 기반 라벨 정제 기법을 통해 pseudo label 노이즈의 영향을 완화할 수 있을 것이다. 또한 pruning과 메모리 검색을 통합적으로 고려한 하이브리드 메모리 관리 전략과 실제 사용자 기반 대화 평가를 포함한 확장 연구를 통해 보다 실질적이고 확장 가능한 장기 대화 메모리 관리 프레임워크로 발전시킬 수 있을 것으로 기대된다.

참고문헌

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... and McGrew, B. (2023), Gpt-4 technical report, ArXiv, abs/2303.08774.

Bae, S., Kwak, D., Kang, S., Lee, M. Y., Kim, S., Jeong, Y., ... and Sung, N. (2022), Keep Me Updated! Memory Management in Long-term Conversations, In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3769-3787.

Bai, A., Jagerman, R., Qin, Z., Yan, L., Kar, P., Lin, B. R., ... and Najork, M. (2023), Regression compatible listwise objectives for calibrated ranking with binary relevance, In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 4502-4508.

Belem, C. G., Pezeshkpour, P., Iso, H., Maekawa, S., Bhutani, N., and Hruschka, E. (2025), From single to multi: How llms hallucinate in multi-document summarization, In *Findings of the Association for Computational Linguistics: NAACL 2025*, 5276-5309.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009), Pearson correlation coefficient, In *Noise reduction in speech processing*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1-4.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... and Amodei, D. (2020), Language models are few-shot learners, *Advances in Neural Information Processing Systems*, **33**, 1877-1901.

Cao, Z., Qin, T., Liu, T. Y., Tsai, M. F., and Li, H. (2007), Learning to rank: from pairwise approach to listwise approach, In *Proceedings of the 24th International Conference on Machine Learning*, 129-136.

Chen, Y. P., Nishida, N., Nakayama, H., and Matsumoto, Y. (2024), Recent Trends in Personalized Dialogue Generation: A Review of Datasets, Methodologies, and Evaluations, In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 13650-13665.

Chhikara, P., Khant, D., Aryan, S., Singh, T., and Yadav, D. (2025), Mem0: Building production-ready ai agents with scalable long-term memory, ArXiv, abs/2504.19413.

Girshick, R. (2015), Fast r-cnn, In *Proceedings of the IEEE international conference on computer vision*, 1440-1448.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T. S. (2017), Neural collaborative filtering, In *Proceedings of the 26th International Conference on World Wide Web*, 173-182.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... and Chen, W. (2022), Lora: Low-rank adaptation of large language models, *ICLR*, **1**(2), 3.

Jang, J., Boo, M., and Kim, H. (2023), Conversation Chronicles: Towards Diverse Temporal and Relational Dynamics in Multi-Session Conversations, In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13584-13606.

Jarvelin, K. and Kekalainen, J. (2002), Cumulated gain-based evaluation of IR techniques, *ACM Transactions on Information Systems (TOIS)*, **20**(4), 422-446.

Jo, E., Jeong, Y., Park, S., Epstein, D. A., and Kim, Y. H. (2024), Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention, In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1-21.

Kang, J., Kim, H., and Kim, H. (2025), Generation-Based and Emotion-Reflected Memory Update: Creating the KEEM Dataset for Better Long-Term Conversation, In *Proceedings of the 31st International Conference on Computational Linguistics*, 9260-9277.

Kasahara, T., Kawahara, D., Tung, N., Li, S., Shinzato, K., and Sato, T. (2022), Building a Personalized Dialogue System with Prompt-Tuning, In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 96-105.

Koren, Y., Bell, R., and Volinsky, C. (2009), Matrix factorization techniques for recommender systems, *Computer*, **42**(8), 30-37.

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., ... and Liu, H. (2025), From generation to judgment: Opportunities and challenges of llm-as-a-judge, In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2757-2791.

Li, H., Yang, C., Zhang, A., Deng, Y., Wang, X., and Chua, T. S. (2025), Hello again! llm-powered personalized agent for long-term dialogue, In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers), 5259-5276.

Lin, C. Y. (2004), Rouge: A package for automatic evaluation of summaries, In *Text Summarization Branches Out*, 74-81.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023), G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment, In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511-2522.

Liu, D., Wu, Z., Song, D., and Huang, H. Y. (2025), A Persona-Aware LLM-Enhanced Framework for Multi-Session Personalized Dialogue Generation, In *Findings of the Association for Computational Linguistics: ACL 2025*, 103-123.

Lu, J., An, S., Lin, M., Pergola, G., He, Y., Yin, D., ... and Wu, Y. (2023), Memochat: Tuning llms to use memos for consistent long-range open-domain conversation, ArXiv, abs/2308.08239.

Maharana, A., Lee, D. H., Tulyakov, S., Bansal, M., Barbieri, F., and Fang, Y. (2024), Evaluating Very Long-Term Conversational Memory of LLM Agents, In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 13851-13870.

Menon, A. K., Jiang, X. J., Vembu, S., Elkan, C., and Ohno-Machado, L. (2012), Predicting accurate probabilities with a ranking loss, In *Proceedings of the... International Conference on Machine Learning. International Conference on Machine Learning* (Vol. 2012), 703.

Ong, K. T. I., Kim, N., Gwak, M., Chae, H., Kwon, T., Jo, Y., ... and Yeo, J. (2025), Towards lifelong dialogue agents via timeline-based memory management, In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers), 8631-8661.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022), Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems*, **35**, 27730-27744.

Pan, Z., Wu, Q., Jiang, H., Luo, X., Cheng, H., Li, D., ... and Gao, J. (2025), Secom: On memory construction and retrieval for personalized conversational agents, In *The Thirteenth International Conference on Learning Representations*.

Papineni, K., Roukos, S., Ward, T., and 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*, 311-318.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. (2021), Mauve: Measuring the gap between neural text and human text using divergence frontiers, *Advances in Neural Information Processing Systems*, **34**, 4816-4828.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023), Direct preference optimization: Your language model is secretly a reward model, *Advances in Neural Information Processing Systems*, **36**, 53728-53741.

Reimers, N. and Gurevych, I. (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)*, 3982-3992.

Sculley, D. (2010), Combined regression and ranking, In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 979-988.

Spearman, C. (1904), The Proof and Measurement of Association between Two Things, The American Journal of Psychology, 15(1), 72-101.

Tan, Z., Yan, J., Hsu, I. H., Han, R., Wang, Z., Le, L., ... and Pfister, T. (2025), In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents, In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8416-8439.

Terven, J., Cordova-Esparza, D. M., Romero-Gonzalez, J. A., Ramirez-Pedraza, A., and Chavez-Urbiola, E. A. (2025), A comprehensive survey of loss functions and metrics in deep learning, *Artificial Intelligence Review*, **58**(7), 195.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... and Lample, G. (2023), Llama: Open and efficient foundation language models, Arxiv, abs/2302.13971.

Wang, H., Wang, L., Du, Y., Chen, L., Zhou, J., Wang, Y., and Wong, K. F. (2023), A survey of the evolution of language model-based dialogue systems, ArXiv, abs/2311.16789.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024), Multilingual e5 text embeddings: A technical report, Arxiv, abs/2402.05672.

Wang, Q., Fu, Y., Cao, Y., Wang, S., Tian, Z., and Ding, L. (2025), Recursively summarizing enables long-term dialogue memory in large language models, *Neurocomputing*, **639**, 130193.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... and Zhou, D. (2022), Chain-of-thought prompting elicits reasoning in large language models, *Advances in Neural Information Processing Systems*, **35**, 24824-24837.

Wilcoxon, F. (1945), Individual comparisons by ranking methods, *Biometrics Bulletin*, **1**(6), 80-83.

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., and Nie, J. Y. (2024), C-pack: Packed resources for general chinese embeddings, In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 641-649.

Xu, J., Szlam, A., and Weston, J. (2022), Beyond goldfish memory: Long-term open-domain conversation, In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (volume 1: long papers), 5180-5197.

Xu, X., Gou, Z., Wu, W., Niu, Z. Y., Wu, H., Wang, H., and Wang, S. (2022), Long Time No See! Open-Domain Conversation with Long-Term Persona Memory, In *Findings of the Association for Computational Linguistics: ACL 2022*, 2639-2650.

Yan, L., Qin, Z., Wang, X., Bendersky, M., and Najork, M. (2022), Scale calibration of deep ranking models, In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4300-4309.

Yi, Z., Ouyang, J., Xu, Z., Liu, Y., Liao, T., Luo, H., and Shen, Y. (2024), A survey on recent advances in llm-based multi-turn dialogue systems, *ACM Computing Surveys*.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and 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* (Volume 1: Long Papers), 2204-2213.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019), Bertscore: Evaluating text generation with bert, ArXiv, abs/1904.09675.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., ... and Zhou, J. (2025), Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models, Arxiv, abs/2506.05176.

Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... and Stoica, I. (2023), Judging llm-as-a-judge with mt-bench and chatbot arena, *Advances in Neural Information Processing Systems*, **36**, 46595-46623.

Zhong, H., Dou, Z., Zhu, Y., Qian, H., and Wen, J. R. (2022), Less is More: Learning to Refine Dialogue History for Personalized Dialogue Generation, In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5808-5820.

Zhong, W., Guo, L., Gao, Q., Ye, H., and Wang, Y. (2024), Memorybank: Enhancing large language models with long-term memory, In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 17), 19724-19731.

<부록>

A. 요약 항목 유용성에 대한 사례 연구

제안한 유용성 정의와 pruning 전략이 대화 맥락에서 어떻게 작동하는지를 정성적으로 분석하기 위해, <Table A1>에 대표적인 사례를 제시한다. 각 사례는 원 대화 일부, 해당 세션에서 생성된 요약 항목, 그리고 이에 대한 유용성 판단 유형(type)을 함께 포함한다. 이를 통해 모델이 어떤 요약 항목을 저유용성으로 판단하여 제거하는지, 그리고 어떤 항목을 이후 대화에서 활용 가능한 정보로 유지하는지를 구체적으로 확인할 수 있다.

Table A1. Case study of Summary Bullet Usefulness

Original Dialogue Turns (Excerpt)	Summary Bullet	Type
S1: I was researching more bands to listen not too long ago, and I found one that I really enjoy that I haven't heard of before. Their music is really good and unique, and something I really recommend you should listen to. I'll send you a link to a YouTube video so you can have a listen. S2: Wow! That would be really awesome. I would totally love that.	S2 is enthusiastic and would love to receive the link.	Low Usefulness (Ambiguous reference)
S2: Its not like I'm an alchoholic like my dad. I do drink from time to time but not to excess like him. S1: That's also part of the fun of being an uncle: I can spend time with the kid and then give them back. Do you have a favorite kind of car to work on?	S2 says “Its not like I’m an alchoholic like my dad. I do drink from time to time but not to excess like him.”	Low Usefulness (Poor abstraction)
S1: Very cool! Did you go to a movie theater to see it or watch it at home? S2: I watch it at home with my girlfriend she also loves horror movies. S1: Haha yeah, I remember you mentioning that. Do you guys watch movies together often?	S2 watched the movie at home with his girlfriend, who also loves horror movies.	High Usefulness (Reusable context)

<Table A1>의 사례들은 요약 항목의 유용성이 개별 발화의 내용 자체보다는, 이후 대화 맥락에서 독립적으로 재참조될 수 있는 정보인지 여부에 의해 좌우됨을 보여준다. 저유용성으로 판단된 요약 항목은 주로 특정 발화에 대한 반응이나 개인적 진술을 그대로 요약한 형태로, 원 대화 맥락이 함께 주어지지 않을 경우 의미가 충분히 전달되지 않거나, 장기적으로 활용 가능한 정보로 일반화되지 못하는 특성을 보인다. 이러한 항목들은 이후 세션에서 재활용 가능성이 낮아 pruning 대상으로 판단된다.

반면, 고유용성으로 평가된 요약 항목은 사용자의 선호, 활동, 관계와 같이 이후 대화에서도 반복적으로 활용될 수 있는 맥락 정보를 간결하게 포착하고 있다. 이와 같은 요약은 단일 발화의 내용을 넘어, 대화 전반에서 의미 있는 상태 정보를 제공함으로써 이후 응답 생성 과정에 직접적으로 기여할 수 있다. 이러한 관찰은 제안한 유용성 기준이 요약의 문장 수준 품질이 아니라, 실제 대화 흐름에서의 활용 가능성을 중심으로 설계되었음을 뒷받침한다.

B. 후보 임베딩 모델별 성능 비교 결과

Table A2. Performance Comparison of Candidate Embedding Models

Metric	E5-Large			Qwen3-0.6B			BGE-Large-En		
	Vanilla	Stage1	Stage2	Vanilla	Stage1	Stage2	Vanilla	Stage1	Stage2
Global Regression									
MSE	0.1970	0.1801	0.1772	0.1930	0.1875	0.1872	0.1935	0.1811	0.1789
MAE	0.4155	0.3981	0.3881	0.4250	0.4197	0.4094	0.4186	0.3952	0.3876
Pearson	0.0424	0.2184	0.2393	0.0163	0.0218	0.0129	0.0528	0.2200	0.2223
Spearman	0.0350	0.2237	0.2325	0.0622	0.0308	0.0049	0.0665	0.2266	0.2268
Session-wise Ranking									
Spearman	-0.0177	0.2308	0.2594	0.0205	0.0508	0.0716	0.0517	0.2264	0.2289
NDCG@5	0.5998	0.7057	0.7079	0.6054	0.6177	0.6201	0.5958	0.7059	0.7073
Hit@5	0.7467	0.8200	0.8333	0.6000	0.5967	0.6000	0.7216	0.8167	0.8300
Precision@5	0.3823	0.4297	0.4332	0.3759	0.3852	0.3859	0.3802	0.4210	0.4311
Recall@5	0.7114	0.8308	0.8363	0.7163	0.7243	0.7279	0.7406	0.8262	0.8246

C. 모델 성능에 대한 통계적 유의성 검정 결과

Table A3. Exact p-values for Statistical Significance Tests between Models

Metric		Test Type	P-value
Global Regression	MSE	Paired t-test	0.0335
	MAE		0.0039
	Pearson		5.0×10^{-4}
	Spearman		1.0×10^{-3}
Session-wise Ranking	Spearman	Wilcoxon signed-rank test	2.08×10^{-14}
	NDCG@5		2.43×10^{-15}
	Hit@5		2.29×10^{-5}
	Precision@5		3.47×10^{-18}
	Recall@5		1.99×10^{-10}

D. K값 변화에 따른 세션 단위 랭킹 성능의 상세 결과

Table A4. Detailed Session-level Ranking Performance of the Stage 2 Model Across Different k Values

Metric	NDCG@k			Hit@k			Precision@k			Recall@k		
	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5
Vanilla	0.4057	0.4774	0.5909	0.1333	0.4367	0.7100	0.3710	0.3722	0.3746	0.1453	0.4146	0.7059
Stage 1	0.5802	0.6105	0.7196	0.2367	0.5333	0.8567	0.5359	0.4735	0.4377	0.2202	0.5605	0.8482
Stage 2	0.6068	0.6294	0.7297	0.2456	0.5800	0.8733	0.5701	0.4876	0.4403	0.2388	0.5730	0.8510

E. 임계치(Threshold) 민감도 분석

본문에서 사용한 pruning threshold 설정의 타당성을 검증하기 위해, threshold 값 변화에 따른 메모리 압축 특성을 분석한다. Threshold가 메모리 효율성과 정보 보존에 미치는 영향을 정량적으로 파악하는 것을 목적으로 하며, 본문 실험에서 threshold 값을 고정하여 사용한 근거를 제공한다.

Threshold는 예측 유용성 점수에 대한 절대적 기준으로, 유용성이 낮은 메모리 항목을 사전에 제거하기 위한 품질 필터로 사용된다. Threshold 값을 0.4부터 0.7까지 단계적으로 변화시키며, 각 설정에서 세션당 평균적으로 유지되는 메모리 항목 수를 측정하였다. Threshold의 영향을 독립적으로 분석하기 위해, 본 분석에서는 top-k 제한을 적용하지 않았다.

(1) Threshold 변화에 따른 메모리 압축 특성

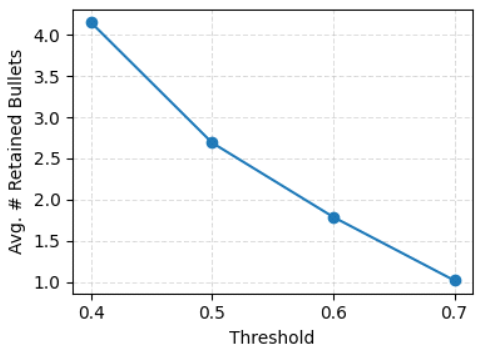

Figure A1. Average Number of Retained Bullets under Different Threshold Settings

<Figure A1>은 threshold 변화에 따른 세션당 평균 유지 메모리 항목 수를 나타낸다. Threshold가 증가함에 따라 유지되는 메모리 항목 수는 전반적으로 감소하는 경향을 보인다. 특히 threshold를 0.4에서 0.5로 증가시킬 때는 비교적 완만한 감소가 관찰되

지만, threshold가 0.6 이상으로 증가할 경우 유지되는 메모리 항목 수가 급격히 감소하는 구간이 나타난다.

이러한 결과는 threshold가 일정 수준을 초과할 경우 pruning 강도가 급격히 증가하여, 중요 메모리 항목까지 함께 제거될 가능성이 높아짐을 시사한다. 반면 threshold = 0.5 설정은 메모리 항목 수를 유의미하게 감소시키면서도 과도한 정보 손실을 유발하지 않는 완충 구간에 해당한다.

(2) Threshold 설정에 대한 논의

Threshold 민감도 분석 결과, threshold는 메모리 압축 강도에 매우 민감하게 작용하는 변수이며, 지나치게 높은 threshold 설정은 과도한 정보 제거로 이어질 수 있음을 확인하였다. 이에 본 연구에서는 메모리 효율성과 정보 보존 간의 균형을 고려하여, 실질적인 메모리 감소 효과를 제공하면서도 지나치게 공격적인 pruning을 방지할 수 있는 threshold = 0.5를 기본 설정으로 선택하였다. 본문에서는 이후 효율성 분석의 인과 해석을 명확히 하기 위해 threshold 값을 고정하고, top-k 변화에 따른 메모리 예산 제어 효과를 중심으로 분석을 수행하였다.

저자소개

이수연: 서울대학교 경영학과에서 2017년 학사, 서울대학교 산업공학과에서 2023년 석사학위를 취득하고 산업공학과 박사과정에 재학 중이다. 연구분야는 Natural Language Processing, Dialogue System, Large Language Models이다.

조성준: 서울대학교 산업공학과에서 학사, 석사학위를 취득하고 미국 워싱턴대학교에서 컴퓨터 사이언스학과에서 인공지능 석사학위 및 메릴랜드대학교 컴퓨터사이언스 학과에서 뉴럴네트워크, 머신러닝 분야로 박사학위를 취득하였다. 이후 공공데이터전략위원장, 정부3.0추진위원회 빅데이터전문위원장과 한국BI데이터마이닝 회장 등을 역임했다. 현재 서울대학교 산업공학과 교수 및 빅데이터 AI 센터장으로 재직하고 있다. 연구분야는 딥러닝, 텍스트마이닝 등 빅데이터 및 AI, 산업 응용이다.