Memory Cost Calculator

The calculator estimates the annual LLM and embedding cost of extracting memories from customer conversations. It is calibrated to the 500-user LongMemEval small split.

What the estimate includes

- LLM input used during memory extraction
- LLM output generated during memory extraction
- `text-embedding-3-small` embedding usage
- Projected memory count and memory-token volume

Query-serving costs, Redis Cloud charges, caching, discounts, reranking, and other infrastructure costs are excluded.

Inputs

Input	Meaning
LLM model	Model and list price used for memory extraction
Active customers	Customers generating conversations
Conversations / customer	Average conversations generated per customer each month
Queries / conversation	Average user queries within one conversation
Average user query	Average user-message length in tokens
Average LLM response	Average assistant-response length in tokens
Memory strategy	Instruct or Remis + Instruct

The customer-support preset uses:

- 1,000,000 active customers
- 0.5 support conversations per customer per month
- 4.2 customer messages per conversation
- 30 tokens per user query
- 120 tokens per LLM response
- GPT-4o
- Instruct strategy

The contact ratio is based on the Corebee benchmark, and conversation depth is based on LoopReply's 10,000-conversation study. Query and response lengths are planning assumptions.

Workload calculation

```
annual queries
  = active customers
  × conversations per customer per month
  × queries per conversation
  × 12

annual source tokens
  = annual queries
  × (average user-query tokens + average LLM-response tokens)
```

All projections scale from this annual source-token volume.

Strategy calibration

The baseline dataset contains:

- 500 users
- 47.8 sessions per user
- 2,125 tokens per session
- 101,575 source tokens per user

The calculator uses these measured average strategy coefficients:

Coefficient	Instruct	Remis + Instruct
Memories / baseline user	414	781
Tokens / memory	25.4	154

Source: LongMemEval strategy worksheet.

Memory projection

Memory count scales with conversation-token volume:

```
source tokens per user per year
  = annual source tokens ÷ active customers

projected memories per user per year
  = source tokens per user per year
  × benchmark memories per user
  ÷ 101,575 baseline source tokens per user

total projected memories
  = projected memories per user per year × active customers

projected memory tokens
  = total projected memories × benchmark tokens per memory
```

Active-customer count changes the total number of memories, but not memories per user when per-user conversation usage remains unchanged.

Ingestion-token calculation

Measured LongMemEval usage is converted into linear scaling ratios.

```
LLM input tokens
  = annual source tokens
  × 95,000,000
  ÷ (500 × 101,575)

LLM output tokens
  = annual source tokens
  × 9,000,000
  ÷ (500 × 101,575)
```

Embedding ratios differ by strategy:

```
Instruct embedding tokens
  = annual source tokens
  × 57,000,000
  ÷ (500 × 101,575)

Remis + Instruct embedding tokens
  = annual source tokens
  × 164,000,000
  ÷ (500 × 101,575)
```

Memory-creation prompt assumption

The 95M input-token measurement includes the creation prompt and other extraction overhead from the benchmark runs. The calculator therefore includes that overhead implicitly through the input-token multiplier.

It does not model fixed prompt tokens per extraction call. LongMemEval averages 2,125 source tokens per session; assuming one extraction call per session, the aggregate measurement implies roughly 1,850 prompt/overhead tokens per call. Workloads with shorter conversations and more extraction calls per source token can cost more than the estimate. Workloads with longer conversations can cost less.

Ingestion-cost calculation

```
LLM input cost
  = LLM input tokens ÷ 1,000,000 × selected model input price

LLM output cost
  = LLM output tokens ÷ 1,000,000 × selected model output price

embedding cost
  = embedding tokens ÷ 1,000,000 × $0.02

annual ingestion cost
  = LLM input cost + LLM output cost + embedding cost
```

Prices are USD list prices last updated in July 2026.

Limitations

- Scaling is linear and assumes similar conversation information density.
- Token counts are treated consistently across model providers.
- One selected model is used for all extraction work.
- Model list prices may change.
- Query-serving costs are intentionally excluded.