B.4 User Utterance Generation Hyperparameters

Table 6: Batching configuration by chat size and domain category for user-turn question generation. NUM_SUBPLANS denotes the number of conversation sub-plans, K the number of batches per sub-plan, and I the number of questions generated per batch.

Chat Size	Category	NUM_SUBPLANS	K	I
128K	General	5	10	2
	Coding	3	23	1
	Math	3	25	1
500K	General	10	10	4
	Coding	10	10	3
	Math	10	10	4
1M	General	10	10	9
	Coding	10	10	6
	Math	10	10	6
10M	General	10	10	9
	Coding	10	10	6
	Math	10	10	6

B.5 Created Probing Questions Distribution

We measure which parts of the dialogue contain the information required to answer the probing questions. To this end, each conversation is divided into ten equal segments, and we record the segment(s) where the supporting evidence for each probing question resides. The detailed methodology for aligning probing questions with dialogue segments is described in Section 2.3. The resulting distributions across conversation lengths are reported in Table 7.

Table 7: Percentage distribution of created probing questions across ten equal chat segments (deciles) for different chat sizes. Each row corresponds to a segment of the dialogue, moving from the beginning (Segment 1) to the end (Segment 10).

Chat Segment (Decile)	100K	500K	1M	10M
1	0.00%	0.65%	0.19%	0.00%
2	11.05%	23.70%	21.60%	10.24%
3	14.83%	15.91%	20.11%	16.27%
4	12.79%	14.45%	15.83%	15.06%
5	13.08%	7.95%	9.50%	14.46%
6	13.37%	9.09%	8.01%	9.64%
7	11.92%	6.33%	5.96%	10.24%
8	8.14%	5.52%	5.21%	13.25%
9	9.59%	4.55%	4.47%	8.43%
10	5.23%	11.85%	9.12%	2.41%

B.6 Memory Abilities Examples

To illustrate how our benchmark evaluates different aspects of long-term conversational memory, we provide representative probing questions and their ideal answers for each of the ten memory abilities. These examples demonstrate how each ability is operationalized in practice.

1. **Abstention (withholding answers when information is missing)**