Research Report: Zeta

Zeta, an Experimental Memory Model:
BEAM-10M, the Reader's Effort, and the Models since July

Sunwoo Kim • Wontopos L.L.C.

WONTOPOS

18 November 2026

Zeta, an Experimental Memory Model: BEAM-10M, the Reader's Effort, and the Models since July

Sunwoo Kim
Wontopos L.L.C.
official@wontopos.com

18 November 2026

Abstract

We report measurements of Zeta, an experimental model that starts from Scroll 2, on BEAM-10M, where one conversation is about ten million tokens. With the same reader (the language model that answers from what the engine hands over, here Claude Opus 5.5) and the same judge (gpt-4.1-mini), Zeta scores 70.39 and 70.21 in two runs against 65.48 for Scroll 2, a paired difference of +4.8 points (95% interval +0.7 to +8.9), and 72.52 in a third run made after one more change to it (+7.0, +2.6 to +11.5). One part of Zeta can be switched off. Having it on is worth +1.1 points (−1.8 to +3.9) when the reader works at maximum effort and searches again about seven times a question, and +3.6 (−0.4 to +7.5) when it works at medium effort and searches again about once; both intervals include zero. What it changes is what the reader is handed first. With the reader at low or medium effort Zeta scores 64.64 and 64.58, within a point of Scroll 2 at maximum effort (−0.9 at medium effort, 95% interval −5.5 to +3.7), for about a tenth of the reader's cost. We also report four earlier models on the same store, a run on BEAM-1M (74.62), how far a repeated run moves, what the event-ordering score leaves uncounted, and where Zeta is weaker than Scroll 2. Zeta is not a released product, and this report does not describe what was changed in it or how it works.

1 Introduction

A memory engine for a language-model application stores what was said and, when asked, hands some of it back. A reader, the language model that answers, sees only what it is handed. Since July 2026 we have measured ours in five reports [Kim, 2026a,b,c,d,e], on conversations that grew from about 140,000 tokens to one million.

This report measures Zeta, an experimental model that starts from Scroll 2, on BEAM-10M [Tavakoli et al., 2025], where one conversation is about ten million tokens. At that length what the reader is handed first (the first list), and what a further search costs, decide much of the answer. One part of Zeta can be switched off, and the runs with it off are reported beside the others.

The findings, all on BEAM-10M with 200 questions unless stated:

1. With the same reader and judge, Zeta scores +4.8 points above Scroll 2 (95% interval +0.7 to +8.9), with fewer tokens of memory and more tool calls, and +7.0 above it in a third run made after one more change (+2.6 to +11.5) (Section 5.1).

2. One part of Zeta can be switched off. What it is worth depends on how much the reader searches again: +1.1 points with the reader at maximum effort and +3.6 at medium effort. Both intervals include zero (Section 5.2).

3. That part changes what the reader is handed first: about 5 of the 14 lines of the first list, and the share of the benchmark's evidence in front of the reader (Section 5.4).

4. With the reader at low, medium, high or xhigh effort Zeta scores 63.4 to 64.8, and at maximum effort 70.3. At medium effort it is within a point of Scroll 2 at maximum effort (−0.9, 95% interval −5.5 to +3.7), for about a tenth of the reader's cost. What maximum effort adds is on the questions about time (Section 5.6).

5. Four earlier models, asked the same questions on the same store, score 52.95 (Tablet 2), 53.09 (Scroll 1.2), 44.65 (Tablet 1) and 38.23 (Scroll 1) (Section 5.7).

6. On BEAM-1M Zeta scores 74.62, with 3,974 tokens of memory per question (Section 5.8).

7. The same run repeated at maximum effort gives totals 0.2 points apart and different scores on 68 of the 200 questions (Section 5.9).

8. The event-ordering score counts order only among the rubric items its line matching paired: 70.6% of the items are in Zeta's answers and 34.1% were paired (Section 5.10).

9. Zeta is weaker than Scroll 2 at saying that the record holds no answer, and on multi-session reasoning (Section 6).

2 Background

2.1 The models since July

The first report, in July [Kim, 2026b], measured on LongMemEval-S [Wu et al., 2025]: 500 questions, each carrying about 140,000 tokens of earlier conversation. The right memory was found 99.6% of the time, and the answers scored 85.2% for Tablet 1 and 90.7% for Scroll 1. The 9 to 14 points between found and answered were what that report was about. A second report that month measured when it pays to hand over more [Kim, 2026a]. Scroll 1.2 scored 92.3% on the same benchmark [Kim, 2026c].

Tablet 2, in August, widened the engine to languages and images, and was the first we measured on BEAM-1M [Kim, 2026e]. Much of that report was about how little one score says on its own: one setting that reports rarely state moved BEAM-1M by 8.9 points. Scroll 2, in November, scored +4.6 points above Tablet 2 per question on BEAM-1M [Kim, 2026d]. That report measured separately what the engine handed over, what the reader used, and how the benchmark scores, and it named weak points: temporal reasoning, event ordering, and saying that the record does not hold the answer. Table 1 lists the models.

Table 1: The models since July. Each row has its own benchmark, reader and scoring, so the scores are not compared down the column. Beside each model is what its own report measured against the one before it. The numbers of the first four rows are those of the earlier reports, unchanged.

Published	Model	Benchmark and score	Against the model before it
7 and 15 July	Tablet 1, Scroll 1	LongMemEval-S 85.2, 90.7	The right memory was found 99.6% of the time
19 July	Scroll 1.2	LongMemEval-S 92.3	+1.6 over Scroll 1. Five runs each; the readers differ, to Scroll 1.2's disadvantage
24 August	Tablet 2	LongMemEval-S 95.7, BEAM-1M 67.5	Read by the same model, Scroll 1.2 scores 92.3 and Tablet 2 93.7. Separate campaigns, so not reported as an established difference
5 November	Scroll 2	BEAM-1M 72.1, and 73.2 with memory tools	+4.6 over Tablet 2 (95% interval +3.4 to +5.9); with memory tools +5.7 (+4.1 to +7.2)
This report	Zeta, experimental	BEAM-10M 70.3 with memory tools	+4.8 over Scroll 2 (95% interval +0.7 to +8.9)

2.2 Related work

Long-term memory for chat agents has been built as a hierarchy the model manages [Packer et al., 2023], as a pipeline that extracts facts and updates or removes them [Chhikara et al., 2025], and as a store with its own updating rule [Zhong et al., 2024]. Generative Agents recall from a stream of what an agent observed [Park et al., 2023]. This report does not compare against these systems. It measures one engine against the model before it, and against itself with one part switched off. LongMemEval [Wu et al., 2025] tests recall over conversations of about 140,000 tokens; BEAM [Tavakoli et al., 2025] extends the setting to one and ten million tokens and ten abilities, and grades each answer against rubric items.

3 Zeta

Zeta is an experimental model that starts from Scroll 2 and differs from it in several ways. One of those differences can be switched off on its own. This report gives the runs with it switched off beside the others and marks them "off". The engine is proprietary: this report measures the result, and does not describe what was changed or how it works. Zeta is not a released product.

4 Method

Benchmark. BEAM-10M has ten conversations of about ten million tokens each and 200 questions, twenty for each of ten abilities. Section 5.8 uses BEAM-1M, 700 questions.

Models. Scroll 2 and the earlier models are measured as they stood on 8 November 2026. Zeta starts from our engine as it stood on 10 November 2026. Scroll 2 is the released model. Every model reads the same stored conversations, and each is used as an application would use it.

Reader and judge. The reader is the language model that answers each question from what the engine hands it: Claude Opus 5.5, called through the Claude Code command line with no tools

other than the memory tools. Effort is that command line's reasoning-effort setting, which has five levels (low, medium, high, xhigh and max); it is maximum unless stated. The judge is gpt-4.1-mini. The reader and the judge are the same for every run.

Memory tools. The application asks once and hands the reader a first list, 14 lines on average for Zeta. Scroll 2 and Zeta come with memory tools: the reader may then look further with them, and decides when to stop. The models before Scroll 2 have no such tools: their reader answers from one reply.

Score. BEAM grades each answer against rubric items, each judged 1, 0.5 or 0, and the question's score is their mean. Event ordering is scored by Kendall's τ_b [Kendall, 1945] between the answered and the reference order, as the benchmark's authors score it. We give that score throughout, and beside it the reading in which event ordering is graded by rubric like the other abilities.

Evidence shown. For 176 of the 200 questions BEAM marks the messages that hold the evidence. Evidence shown is the share of those messages that were in anything the reader received, averaged over questions. No language model is involved.

Memory and cost. Tokens of memory count the memory text the reader received for one question. The reader's cost is each question's usage record at Claude Opus 5.5's list price, as the command-line reader counted it; the runs themselves were answered on a subscription.

Paired comparison. Runs are compared per question. We report the mean difference and a 95% confidence interval, that mean plus or minus 1.96 standard errors; "interval" below means this. Where a condition was run twice, the mean of its two runs is used for each question. Table 2 lists the runs.

Table 2: The runs of this report. All on BEAM-10M, 200 questions, except the last. "Off" is a run with one part of Zeta switched off (Section 3). The third run was made after one more change to Zeta (Section 5.1).

Model	Reader effort	Runs
Scroll 2	maximum, medium	One each
Zeta	maximum	Two
Zeta	low, medium, high, xhigh	One each
Zeta, off	maximum, medium	One each
Zeta, third run	maximum	One
Tablet 2, Scroll 1.2, Tablet 1, Scroll 1	maximum	One each
Zeta, on BEAM-1M	maximum	One

5 Results

5.1 Zeta and Scroll 2

Zeta scored 70.39 and 70.21, the same settings run twice. Scroll 2, with the same reader and the same judge, scored 65.48. The difference per question is +4.8 points (95% interval +0.7 to +8.9). With event ordering graded by rubric as well, Zeta's score is 73.9 and Scroll 2's 67.4 (Table 3, Figure 1; by ability in Table 4). The reader received 5,444 tokens of memory per question on

average, fewer than the 5,794 it received from Scroll 2, and it was shown 66% of the evidence the benchmark marks, against 47%. Zeta's configuration was chosen on this benchmark (Section 6).

After these two runs one more change was made to Zeta. The third run, made after it, scores 72.52: +7.0 points above Scroll 2 (95% interval +2.6 to +11.5) and +2.2 above its two earlier runs (−1.2 to +5.7). The difference is concentrated in the 60 questions of the three abilities this report groups as asking about time (event ordering, summarization and temporal reasoning), which go from 53.5 to 60.7 (+7.2, 95% interval −0.3 to +14.8); the other 140 questions stay where they were (77.5, then 77.6). The third run was made in two parts, those 60 questions and then the other 140. The runs with the part off and the runs at lower effort (Sections 5.2 to 5.6) were made before this change.

Table 3: BEAM-10M, 200 questions. Reader Claude Opus 5.5 at maximum effort, judge gpt-4.1-mini. Evidence shown and the questions where none of it was shown are over the 176 questions that mark evidence. "Off" is a run with one part of Zeta switched off (Section 3). The third run was made after one more change to Zeta.

	Score (BEAM's scoring)	Score, ordering by rubric	Tokens of memory	Tool calls	Evidence shown	None shown
Scroll 2	65.48	67.42	5,794	4.99	47.0%	54
Zeta, off	69.25	72.79	4,835	6.98	58.6%	31
Zeta, first run	70.39	74.01	5,556	7.12	65.1%	20
Zeta, second run	70.21	73.73	5,332	6.87	66.9%	15
Zeta, third run	72.52	76.32	5,031	6.51	67.4%	14

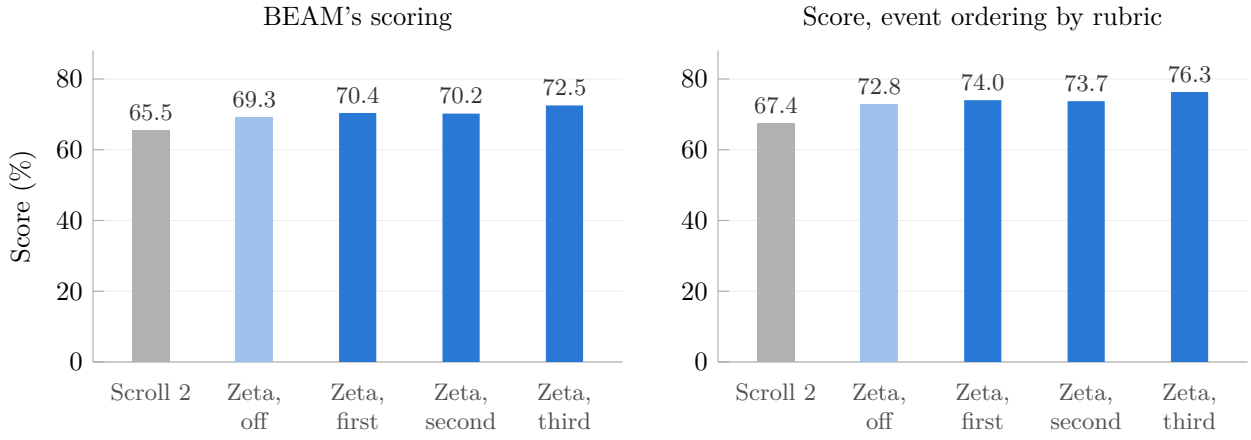


Figure 1: BEAM-10M with the same reader and judge. Paired per question against Scroll 2, the mean of Zeta's two runs is +4.8 points by BEAM's scoring (95% interval +0.7 to +8.9) and +6.5 with event ordering graded by rubric (+2.1 to +10.8).

Table 4: BEAM-10M by ability, twenty questions each, BEAM's scoring. Section 5.9 shows how far one ability moves between two runs of the same settings.

	Scroll 2	Zeta, off	Zeta, first run	Zeta, second run
Abstention	95.0	85.0	77.5	85.0
Contradiction resolution	75.0	71.2	81.2	77.5
Event ordering	25.4	35.5	34.4	36.2
Information extraction	80.0	77.5	75.0	80.0
Instruction following	83.8	91.2	98.8	98.8
Knowledge update	85.0	85.0	82.5	82.5
Multi-session reasoning	45.0	40.4	39.6	37.2
Preference following	71.2	91.2	90.8	78.8
Summarization	53.1	52.8	61.6	57.3
Temporal reasoning	41.2	62.5	62.5	68.8
All 200	65.48	69.25	70.39	70.21
The three about time (60)	39.9	50.3	52.8	54.1

5.2 One part switched off

We switched one part of Zeta off and ran again, in two conditions (Table 5). With the reader at maximum effort, having it on is worth +1.1 points. With the reader at medium effort it is worth +3.6. Both intervals include zero, the second only narrowly.

The two conditions differ in how much the reader searches again. At maximum effort it calls the tools about seven times for a question, so what came up first decides less of the score. At medium effort it calls them about once.

Table 5: Zeta with one part on and off on BEAM-10M, paired per question. At maximum effort the run with it on is the mean of two runs.

Condition	On	Off	Difference	95% interval
Reader at maximum effort	70.30	69.25	+1.05	−1.8 to +3.9
Reader at medium effort	64.58	61.00	+3.58	−0.4 to +7.5

5.3 The rest of the difference

That part is not the only difference between Zeta and Scroll 2. With it off and the reader at maximum effort, Zeta scores 69.25, +3.8 points above Scroll 2 (95% interval −1.1 to +8.7). With it off and the reader at medium effort, Zeta scores 61.00 and Scroll 2 60.11 (+0.9, 95% interval −4.6 to +6.4). Both intervals include zero. This remainder comes from the other differences in Zeta, which this report does not describe.

5.4 What changes when it is on

What the reader is handed first changes (Figure 2). The first list has 14 lines on average, and with the part on, 5.4 of them on average are different lines. Over the 200 questions, 19 of the lines that left were lines of evidence, and 39 of the lines that came in were. The first list holds about 51% of the marked evidence with the part on and 44% with it off.

With the reader at maximum effort, the share of evidence it was shown over the whole question rose from 58.6% to between 65% and 67%, and the questions where it saw none fell from 31 to between 15 and 20. Instruction following is 91.2 with the part off and 98.8 with it on, in both runs (+7.5, 95% interval −0.3 to +15.3, which includes zero; twenty questions).

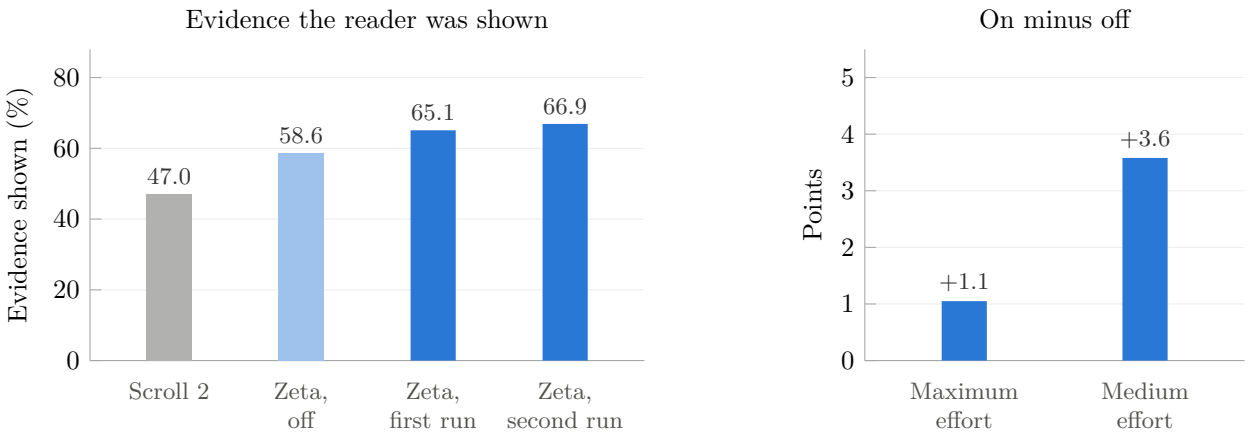


Figure 2: What changes with one part of Zeta on. Left: the share of marked evidence the reader was shown at maximum effort, over the 176 questions that mark evidence. Right: the difference in score between that part on and off in the two conditions of Table 5. Both intervals include zero.

5.5 Cost and coverage

With the reader at maximum effort the total barely moves. Having the part on also costs about 600 more tokens of memory per question (4,835 to 5,444).

In the BEAM-10M runs the part was only partly in effect: for about 30% of what it applies to, and fully in two of the ten conversations. On the 40 questions of those two conversations, at maximum effort, the runs with it on scored 64.0 and the run with it off scored 67.9 (−3.8, 95% interval −8.6 to +1.0). On the other eight conversations the difference is +2.3 (−1.0 to +5.5). On BEAM-10M we have not measured it on and off while fully in effect.

5.6 The reader's effort

The two runs of Section 5.1 used the reader at maximum effort. With the same Zeta, we lowered only the reader's effort and ran four more times, once at each lower level (Table 6, Figure 3). At low, medium, high and xhigh effort the scores are 64.64, 64.58, 64.77 and 63.40. The two runs at maximum effort are +6.0 points above the mean of these four (95% interval +2.6 to +9.3). Raising the effort did not raise the score until the maximum. Each of these runs was made once.

At medium effort the reader searches again about once: it calls the tools 1.25 times per question and receives 1,386 tokens of memory. It gains little from that search: it is shown 53% of the evidence the benchmark marks, and the first list already holds about 51% of it. The score at medium effort differs from Scroll 2's at maximum effort by −0.9 (95% interval −5.5 to +3.7). The reader costs $0.035 per question against $0.332, and takes 10 seconds against 96. At low effort it calls the tools 0.69 times, reads 1,092 tokens of memory and scores the same, for $0.029 per question.

What maximum effort buys is on the questions about time. On the 140 questions that are not about time, medium effort scores 76.0 and maximum effort 77.5 (the interval of the difference, −3.7 to +6.9, includes zero). On the 60 questions about time, medium effort scores 38.0 and maximum

effort 53.5 (+15.4, 95% interval +7.9 to +23.0). Those 60 questions are also most of what maximum effort costs: $0.70 per question against $0.21 for the others.

Scroll 2 itself, with the reader at medium effort, scores 60.11, with 32% of the marked evidence shown. Its first list holds 26% to 28% of that evidence in its two runs. At medium effort Zeta is +4.5 points above Scroll 2 (95% interval −1.1 to +10.1), close to the difference at maximum effort, with an interval that includes zero. Lowering the reader from maximum to medium effort costs both about the same: 5.4 points for Scroll 2 and 5.7 for Zeta.

Table 6: The reader's effort on BEAM-10M, 200 questions. Cost, seconds, tokens and tool calls are per question. Evidence shown is over the 176 questions that mark evidence. "Off" is a run with one part of Zeta switched off (Section 3). The third run was made after one more change to Zeta (Section 5.1).

	Score	Cost	Seconds	Tokens of memory	Tool calls	Output tokens (of which thinking)	Evidence shown
Zeta, low	64.64	$0.029	10	1,092	0.69	403 (89)	52.2%
Zeta, medium	64.58	$0.035	10	1,386	1.25	559 (189)	53.0%
Zeta, high	64.77	$0.043	11	1,618	1.67	674 (257)	54.8%
Zeta, xhigh	63.40	$0.062	16	2,269	2.74	1,116 (598)	56.8%
Zeta, max, first	70.39	$0.353	103	5,556	7.12	10,519 (9,639)	65.1%
Zeta, max, second	70.21	$0.341	102	5,332	6.87	10,181 (9,312)	66.9%
Zeta, max, third	72.52	$0.340	101	5,031	6.51	10,268 (9,389)	67.4%
Zeta, max, off	69.25	$0.339	101	4,835	6.98	10,279 (9,419)	58.6%
Zeta, medium, off	61.00	$0.036	19	1,352	1.46	545 (175)	48.4%
Scroll 2, medium	60.11	$0.041	10	1,908	1.25	567 (176)	32.3%
Scroll 2, max	65.48	$0.332	96	5,794	4.99	9,712 (8,970)	47.0%
Tablet 2, max	52.95	$0.132	42	2,679	0	4,658 (4,408)	33.3%
Scroll 1.2, max	53.09	$0.083	26	1,872	0	2,739 (2,500)	26.1%
Tablet 1, max	44.65	$0.041	16	387	0	1,617 (1,439)	21.1%
Scroll 1, max	38.23	$0.032	12	337	0	1,179 (1,020)	13.9%

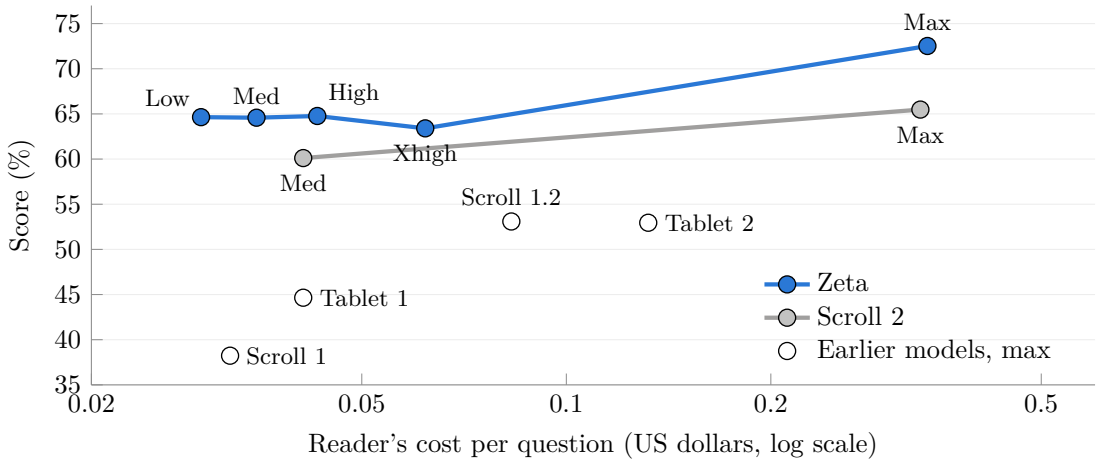


Figure 3: Score against the reader's cost on BEAM-10M, one point for each reader effort. Zeta's point at maximum effort is its third run (72.52); its first two runs at maximum effort score 70.39 and 70.21, and its points at lower effort were measured before the change the third run carries. The grey line is Scroll 2 at medium and at maximum effort. The hollow points are the earlier models at maximum effort. The vertical axis starts at 35.

5.7 Earlier models on the same store

We also asked the four models before Scroll 2 the same 200 questions with the reader at maximum effort. They read the same stored conversations. They have no memory tools, so each was asked in its own configuration and the reader answered from what came back, without searching again. Their answer prompt lacks one sentence that the Scroll 2 and Zeta runs carry, which tells the reader to combine the memories it has; its effect on these models was not measured. Tablet 2 scores 52.95 with 2,679 tokens of memory, Scroll 1.2 53.09 with 1,872, Tablet 1 44.65 with 387, and Scroll 1 38.23 with 337. The share of the marked evidence in front of the reader is 33%, 26%, 21% and 14%.

These numbers are not differences under the same conditions as the rest of this report. Scroll 2 at maximum effort is +12.5 points above Tablet 2 (95% interval +7.6 to +17.4), but one reader could search again and the other could not. Scroll 1 and Scroll 1.2 were measured as engines alone, without the hosted service they were sold with, so their numbers may be below what the products would have scored.

5.8 BEAM-1M

We ran Zeta once on BEAM-1M, 700 questions, as it stood in the third run, with the same reader at maximum effort and the same judge. In this run the part that can be switched off (Section 5.2) was fully in effect. Zeta scores 74.62 by BEAM's scoring and 79.31 with event ordering graded by rubric, with 3,974 tokens of memory per question and 5.75 tool calls.

The Scroll 2 report gives 73.2 for Scroll 2 with memory tools on this benchmark, read by Claude Opus 5 [Kim, 2026d]. The readers differ, so the 1.4 points between the two are not a difference under the same conditions. We have no run of Scroll 2 on BEAM-1M with this reader and memory tools, so this run does not tell whether the difference seen on BEAM-10M holds at this size.

5.9 Repeating a run

We ran the same model with the same settings and the same reader twice. The totals are 70.39 and 70.21, 0.2 points apart. By question, 68 of the 200 scored differently, 39 higher in the second run and 29 lower. One ability is twenty questions, so it moves further: preference following was 90.8 and 78.8, temporal reasoning 62.5 and 68.8, abstention 77.5 and 85.0. So this report reads a difference in one ability only where both runs point the same way.

5.10 What the score does not count

The Tablet 2 report said the event-ordering score measures something other than ordering [Kim, 2026e]. This time we counted it. The benchmark cuts an answer into lines, asks of each line whether it is the same thing as a rubric item, pairs them, and then looks at the order. Graded item by item, 70.6% of the rubric items are present in Zeta's answers. Line matching paired 34.1% of the rubric items. Among the paired items the order scores 89.7 by the benchmark's measure (Table 7). A rubric item is a short name; a line of the answer is a sentence with a date and what happened.

Table 7: Event ordering on BEAM-10M, twenty questions, counted three ways beside the benchmark's score.

	Scroll 2	Zeta, first run
Rubric items present in the answer	44.8%	70.6%
Rubric items paired by line matching	17.8%	34.1%
Order among the paired items	65.2	89.7
Event ordering, BEAM's scoring	25.4	34.4

In temporal reasoning, in Zeta's first run, seven questions had the number of days right and both dates written in the answer, and lost the rubric item that asks for the range from one date to the other. That is 1.4 points of the total. We did not shape the answers to the rubric. The instruction given to the reader is the one used to measure Scroll 2.

6 Limitations

Saying the record holds no answer. A weak point the Scroll 2 report named is still there. On abstention, Scroll 2 scores 95.0 and Zeta 77.5 and 85.0. One reading is that when more is found, the reader builds an answer from it even where the record holds none; we did not measure this.

Multi-session reasoning. Zeta scores below Scroll 2 on multi-session reasoning: 39.6 and 37.2 in its two runs, against 45.0.

Heavy questions. For an event-ordering question the reader receives 11,912 tokens on average and calls the tools 13 times. At maximum effort the reader costs $0.35 and $0.34 per question in Zeta's two runs, against $0.33 with Scroll 2.

Coverage. On BEAM-10M the part of Zeta that was switched on and off was only partly in effect (Section 5.5).

A configuration chosen on this benchmark. Zeta's configuration was chosen on BEAM-10M. BEAM-1M was measured afterwards with the same configuration (Section 5.8).

Runs. Zeta was run twice with the same settings, and every other condition once.

One benchmark, one reader, one language. The new measurements in this report use one family of benchmarks and one reader. BEAM is in English, and scores in other languages were not measured. The score rests on one language-model judge.

The earlier numbers. The figures for Tablet 1, Scroll 1, Scroll 1.2, Tablet 2 and Scroll 2 on LongMemEval-S and BEAM-1M are those of the five earlier reports, unchanged, and each carries the limits its own report states.

An experimental model. Zeta is an experimental model. It is not a released product.

Who wrote this. This is our own product and this report was written by the person who builds it. Every number in it is self-reported.

7 Conclusion

On BEAM-10M, with the same reader and judge, Zeta scores +4.8 points above Scroll 2 with fewer tokens of memory, and +7.0 above it in a third run made after one more change; its configuration was chosen on this benchmark. Having the part of Zeta that can be switched off on adds +1.1 points at maximum effort and +3.6 at medium effort, and neither interval excludes zero. It changes what the reader is handed first: more than a third of the first list, and more of the evidence in front of the reader.

How much that is worth depends on the reader. For a reader that searches again about seven times a question, the part moves the score by about one point. For a reader that searches again about once it is +3.6 points, at about a tenth of the reader's cost. On BEAM-10M the part was only partly in effect, and on the two conversations where it was fully in effect the difference was −3.8 (95% interval −8.6 to +1.0). The runs at lower effort were made once each.

Zeta is still weaker than Scroll 2 at saying that the record holds no answer and on multi-session reasoning.

License. BEAM is released under CC BY-SA 4.0.

References

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. URL `https://arxiv.org/abs/2504.19413`.

M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. doi: 10.1093/biomet/33.3.239.

Sunwoo Kim. When to Deliver More: A Memory-Size Crossover Between Tablet 1 and Scroll 1, 2026a. URL https://wontopos.com/research/crossover. Wontopos research report, 15 July 2026.

Sunwoo Kim. What Remains After Retrieval Saturates: Measuring the Delivery Gap in AI Memory Systems, 2026b. URL https://wontopos.com/research/delivery-gap. Wontopos research report, 7 July 2026.

Sunwoo Kim. Scroll 1.2 on LongMemEval-S: A Five-Run Benchmark Against Scroll 1, 2026c. URL https://wontopos.com/research/scroll-1-2. Wontopos research report, 19 July 2026.

Sunwoo Kim. Wontopos Scroll 2: Score, Evidence Recall and Coverage for a Long-Term Memory Engine, 2026d. URL https://wontopos.com/research/scroll-2. Wontopos research report, 5 November 2026.

Sunwoo Kim. Wontopos Tablet 2: Measuring Multilingual and Multimodal Memory Retrieval Without Lexical Matching. *arXiv preprint arXiv:2608.23920*, 2026e. URL https://arxiv.org/abs/2608.23920.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023. arXiv:2304.03442.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J. Ross Mitchell. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms. *arXiv preprint arXiv:2510.27246*, 2025. URL https://arxiv.org/abs/2510.27246. v1 31 October 2025, v2 21 February 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.10813. arXiv:2410.10813, October 2024.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19724–19731, 2024. arXiv:2305.10250.