Eleven *words*

Reply length in AI companions: what brevity buys, what it costs, how long a reply should be, and how to measure it fairly

MITWA LAB, MUMBAI

IN PLAIN WORDS

Bloom used to answer in about 54 words, and it now answers in about 11. We wanted to know whether that is actually an improvement. Mostly it is, though with conditions. A 54-word reply takes around 14 seconds to read, and if you had written 34 words yourself, the companion ends up with most of the conversation. At 11 words the reply takes 3 seconds and you keep most of it. The catch is that a good length depends on how much there is to say. In the simple model we work through here, a reply that is too short loses more than one that is too long by the same number of words, which means a companion should average 11 words without answering everything in 11. We also explain why AI judges tend to mark short replies down.

ABSTRACT

Reply length is one of the first things people notice about a conversational model and one of the least studied. Spoken turns last about two seconds, while model replies often run to paragraphs. We define a companion's floor share, the fraction of the words in an exchange that belong to it, and show that shortening a reply from 54 words to 11 against a 34-word message takes that share from 61% to 24% and reading time from 13.6 seconds to 2.8. We then ask how long a reply ought to be. In a model where value saturates and every word has a cost, the ideal length is $\tau \ln(B/\kappa\tau)$, which grows with how much there is to say and only logarithmically with what is at stake. In the same model an undershoot costs more than an overshoot of equal size, so aiming at a fixed short length is a mistake, and length should respond to the person's message with an elasticity between zero and one. Because model judges prefer longer text, their scores cannot separate quality from length unless length is controlled for, and we describe a control. Last, we report that the average reply of Bloom, an on-device companion, fell from 54 words to 11, alongside a small internal pilot in which short replies scored well on engagement and poorly on structured technique. We list what an evaluation would have to report before either figure could be called progress. How Bloom was trained is outside the scope of this note.

01 Introduction

Tell a friend that your day went badly and you will probably get something like "That's rough. What happened?" Put the same sentence to a general-purpose assistant and you are likely to get several paragraphs, a list of suggestions and an offer to help further. There is more in the second answer, but by the time you have read it the conversation is no longer really yours.

Ordinary conversation runs on short turns. Measured across many languages, a turn lasts about two seconds on average and the gap before the next speaker starts is around 200 milliseconds (Levinson, 2016; Levinson and Torreira, 2015). Written text works differently, and since language models learned mostly from written text, they tend to answer like writers. For a companion that is a poor fit, yet we have found very little work that asks how much a companion should say in a turn.

This note started from one number. After training, Bloom's average reply went from 54 words to 11, and we wanted to understand what that change does and whether it is a good one. We ended up with four questions: what a shorter reply gives the person, what it takes away, how long a reply should be in the first place, and how to tell a good short reply from one that is merely short. The note makes the following contributions.

- We describe reply length in terms a person can feel: the companion's share of the floor and the time a reply takes to read (Sections 3 and 4).
- We give a model of ideal length and prove two things about it. Length should follow how much there is to say rather than how much is at stake, and a reply that is too short loses more than one that is equally too long (Section 5).
- We set out four costs of brevity: lost content, a weaker impression of care, mismatch with long messages, and the temptation to be brief by asking questions (Section 6).
- We show that scores from model judges mix quality with length, and give a control that separates the two (Section 7).
- We report the Bloom result together with a small pilot that partly contradicts it, and say what a proper evaluation of brevity should contain (Sections 8 to 10).

Bloom is a companion and makes no claim to be a therapist. It runs on the iPhone, with the model, the memory and every reply kept on the phone, so reply length is also a question of how much work the device has to do (Section 4.3).

02 Related work

Conversation. Studies of turn-taking report short turns and quick transitions in every language examined (Levinson and Torreira, 2015; Levinson, 2016). Grice's maxim of quantity says a speaker should be as informative as the exchange needs and no more (Grice, 1975). People talking to each other also drift toward each other's style, down to the number of words per turn; the authors who measured this took it as a sign of engagement and found it unrelated to how well the conversation was rated (Niederhoffer and Pennebaker, 2002).

Length and perceived quality. Raters tend to like longer answers. In one widely cited comparison, chatbot replies to patient questions averaged 211 words and physician replies 52; evaluators preferred the chatbot in 78.6% of cases and rated its replies empathetic or very empathetic 45.1% of the time, against 4.6% for the physicians (Ayers et al., 2023). The study did not vary length independently of who wrote the reply, so it cannot tell us how much of that gap is length. With model judges the effect has been measured directly: GPT-4 favoured longer answers more strongly than human raters did (Saito et al., 2023). Section 7 treats this as a measurement problem.

Listening. Motivational interviewing, a counselling method, has training standards that count a practitioner's reflections against their questions. One reflection per question is the minimum and two is considered good (Moyers, Manuel and Ernst, 2014). We borrow the ratio because it is an established way to quantify listening; we are not making a clinical claim by using it. The approach here follows our earlier note on memory, which dealt with a similar decision about when a companion should hold back (Mitwa Lab, 2026).

03 Formulation

Take a single exchange: the person writes a message of u words and the companion replies with ℓ words. We use two quantities to describe what the length of that reply does. The first we call the companion's *floor share*,

$$\sigma = \frac{\ell}{\ell + u} \tag{1}$$

that is, the fraction of the words in the exchange that came from the companion. The second is reading time. For a reader who manages w words a minute,

$$t = \frac{60\,\ell}{w} \text{ seconds} \tag{2}$$

Throughout we use $w = 238$, the adult average for silent reading of English non-fiction in a meta-analysis of 190 studies (Brysbaert, 2019). Counting words treats them as interchangeable, which they are not, so Table 1 separates the jobs that words in a reply can be doing.

MOVE	EXAMPLE	LOST IF CUT
Acknowledge	"That sounds hard."	Warmth
Reflect	"So the toast is what's keeping you up."	Evidence of listening
Ask	"Is it the speech, or the crowd?"	The next turn
Inform	"Writing it out once tends to help."	Substance
Pad	"It's important to remember that everyone is different."	Nothing

Table 1. Jobs that the words in a reply do, with invented examples. Shortening a reply helps when it is the last row that gets cut.

04 What brevity buys

4.1 The floor

Whoever is producing the words holds the floor. Equation (1) says that a 54-word reply to a 34-word message gives the companion 61% of the exchange, while an 11-word reply to the same message gives it 24% (Figure 1). Shortening the reply therefore does more than save time, because it changes which of the two is doing most of the talking. Since a companion exists for the person's side of the conversation, we regard this as the main benefit.

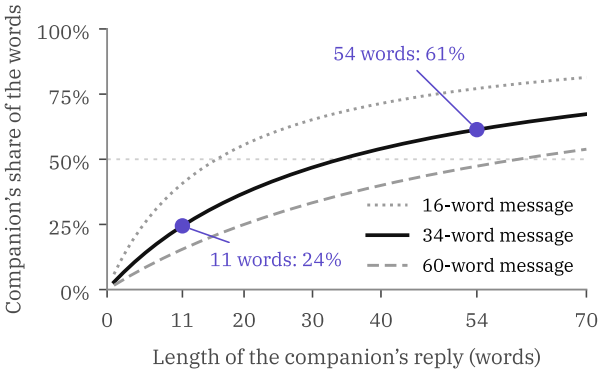


Figure 1. The companion's share of the words in an exchange, from (1), for three message lengths. The dashed line is an even split. The two points are Bloom's average reply before and after, set against a 34-word message.

4.2 Time

From (2), reading a 54-word reply takes 13.6 seconds and reading an 11-word reply takes 2.8. Spoken aloud at 183 words a minute (Brysbaert, 2019) they would last 17.7 and 3.6 seconds. Next to the two-second turns of ordinary speech (Levinson, 2016), a 54-word reply is closer to a short speech than to a turn. To give a sense of scale, thirty replies a day would add up to roughly five and a half extra minutes of reading at the longer length. We offer that only as an illustration, because we have no way of seeing how many replies Bloom's users actually read.

4.3 The device

Bloom produces each reply on the phone, a token at a time, so to a first approximation the time and energy spent on a reply scale with its length. A reply one fifth as long needs about one fifth of the generation work. This follows from how such models generate text; we have not measured it on a device for this note.

05 How long a reply should be

None of this tells us that shorter is always better, so we need some account of what the right length is. Suppose a reply of ℓ words is worth

$$V(\ell) = B(1 - e^{-\ell/\tau}) - \kappa\ell \tag{3}$$

to the person. The first term levels off: early words carry most of the value, and τ is the number of words it takes to deliver about 63% of it. The second term charges κ per word for the reader's time and attention and for the floor the reply occupies. These are assumptions. We picked them because they are the simplest forms that give diminishing returns and a steady cost.

> **Proposition 1 (Ideal length).**
> *If $B > \kappa\tau$, the value (3) is greatest at $\ell^* = \tau \ln(B/\kappa\tau)$. Doubling B adds $\tau \ln 2$, about 0.69τ, words.*

$$\ell^* = \tau \ln\left(\frac{B}{\kappa\tau}\right) \tag{4}$$

The practical reading is that a companion should not write more simply because a moment is important. Doubling the stakes adds a handful of words, whereas the reply grows much faster with the amount there is to say, as (5) below shows. In this model care has to show in the choice of words, since it barely shows in their number.

> **Proposition 2 (Undershooting costs more).**
> *For any x with $0 < x \le \ell^*$, a reply x words shorter than ℓ^* loses more value than a reply x words longer. The difference is $2\kappa[\tau \sinh(x/\tau) - x]$, which is positive and grows with x.*

Figure 2 plots the two losses. They differ because of the shape of (3): the words removed from a too-short reply are early, valuable ones, while the words added to a too-long reply are late and cheap. This result sits awkwardly with our own headline number. A companion trained toward a fixed short length will undershoot whenever the person has brought a lot, and that is the more expensive direction to be wrong in.

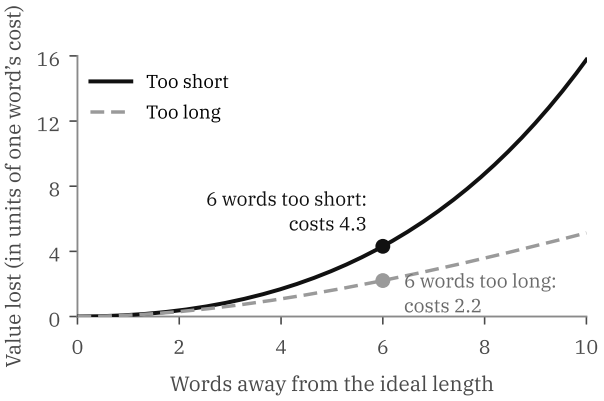


Figure 2. Value lost by missing the ideal length, from (3), with $\tau = 6$ words and the ideal at 11. We chose these parameters to put the ideal at 11 words; the asymmetry itself holds for any parameters.

The model also predicts how length should respond as content grows. Differentiating (4) gives

$$\eta^{*} = \frac{d \ln \ell^{*}}{d \ln \tau} = 1 - \frac{\tau}{\ell^{*}} \tag{5}$$

which is between zero and one as long as the ideal reply is longer than τ. So the reply should lengthen when there is more to say, though less than proportionally. For the illustrative values in Figure 2 the ideal elasticity is 0.45, and for a fixed-length reply it is zero by definition. What we should be aiming at, then, is a relationship between what the person brings and what comes back. An average length, taken alone, cannot tell us whether that relationship is there.

06 What brevity costs

6.1 Content

Let the content of a reply be $c = d\ell$, with d the amount each word carries. Moving from 54 words to 11 divides ℓ by 4.9, so content survives only if density goes up by the same factor:

$$\frac{d_{\text{after}}}{d_{\text{before}}} = \frac{\ell_{\text{before}}}{\ell_{\text{after}}} = \frac{54}{11} \approx 4.9 \tag{6}$$

That can happen if most of the original 54 words were padding in the sense of Table 1. If they were not, something was lost. A word count alone cannot settle which, since the answer depends on what was cut.

6.2 The appearance of care

Readers often take length as a sign of effort. In the chatbot and physician comparison it was the longer replies that were judged empathetic (Ayers et al., 2023). An eleven-word reply forgoes that signal and has to be judged caring on other grounds, chiefly by being specific to the person. We count this as a real cost, and we expect it to matter most early on, before a companion has had any chance to show what it remembers.

6.3 Mismatch

People who are talking tend to match each other's length (Niederhoffer and Pennebaker, 2002). A companion that sends back eleven words whatever it receives will be about right for a short message and far too brief for a long one, and by Proposition 2 the long message is where the loss is greatest. The three pilot conversations in Section 8 show the same pattern.

6.4 Brevity bought with questions

The easiest way to be brief is to ask a question. A question is short, sounds interested and leaves the work with the other person. Suppose a fraction f_{q} of replies are questions. Even if every remaining reply were a reflection, the ratio of reflections to questions could be no higher than

$$\frac{1 - f_{\text{q}}}{f_{\text{q}}} \tag{7}$$

When we analysed 16,027 synthetic conversations written in the short style, the assistant asked a question in 73% of its turns. From (7), its ratio is at most 0.37, well under the one-to-one minimum in the listening standard (Moyers, Manuel and Ernst, 2014). These conversations were synthetic and are not Bloom's replies to real people, so we read the figure as a caution about the style and do not treat it as a measurement of the product.

07 Measuring brevity fairly

Conversational models are now mostly evaluated by other models, and those judges favour length (Saito et al., 2023). To see what that does to a brevity result, write a judge's score for a reply as

$$S = Q + \beta \ln \ell + \varepsilon \tag{8}$$

with Q the quality we actually care about, β the judge's preference for length and ε noise.

> **Proposition 3 (Length-confounded scores).** *Under (8), the expected score gap between two systems A and B is the quality gap plus $\beta \ln(\ell_B/\ell_A)$. The quality gap is not identified from scores alone. It is identified once β is estimated from pairs of replies with the same content and different lengths.*

The reason is short. Subtracting (8) for the two systems leaves one equation in two unknowns. For a

pair with the same content, Q cancels and only the length term remains, which gives

$$\hat{\beta} = \frac{\overline{S_1 - S_2}}{\ln(\ell_1/\ell_2)} \tag{9}$$

In Bloom's case $\ln(11/54) = -1.59$. If a judge had $\beta = 0.5$ on a ten-point scale, the shorter build would lose 0.8 points with no change in quality at all. We have not estimated β for any of the judges we used, so that figure is only an illustration. What it shows is that a lower score after shortening may or may not mean lower quality, and equally that an unchanged score may hide an improvement. Either way the judge's preference for length has to be measured first, which is what (9) does.

Adaptation needs its own measure. Over a set of exchanges the observed elasticity is

$$\eta = \frac{\mathrm{Cov}(\ln \ell,\ \ln u)}{\mathrm{Var}(\ln u)} \tag{10}$$

and it should be compared with the ideal value in (5). Mean length describes how much a companion talks; the elasticity describes whether the amount it says depends on the amount it was told.

08 Result

We put two questions to the data: whether Bloom's replies became shorter, and whether there is any evidence that the shorter replies are better. On the first, the average reply went from 54 words to 11 after training. Table 2 lists what follows from those two figures.

	BEFORE	AFTER
Average reply	54 words	11 words
Time to read	13.6 s	2.8 s
Time if spoken	17.7 s	3.6 s
Share of words, 34-word message	61%	24%
Generation work, relative	1	about 0.2

Table 2. Bloom's average reply length, and what (1) and (2) give for it. The reply is 80% shorter, a factor of 4.9.

The second question is harder, and what evidence we have is mixed. In an internal pilot, model judges scored three synthetic conversations written in the short style (Table 3).

	MESSAGE	REPLY	SHARE	ENGAGE.	TECHN.
C1	25.3	8.4	25%	9.09	4.27
C2	16.3	10.5	39%	9.27	5.27
C3	60.0	12.0	17%	7.36	4.00

Table 3. Internal pilot, three synthetic conversations. Message and reply are average words per turn; share is the assistant's floor share from (1); scores are out of 10 from model judges. Both sides were generated, and none of this is Bloom replying to people.

Engagement scores were high and scores for structured technique were low; the short replies were pleasant to read but did little work. Conversation C3 was weakest on both scores. It also had the longest messages and the lowest floor share, which is the mismatch described in Section 6.3. Across the three conversations, average reply length moved only from 8 to 12 words while average message length ran from 16 to 60, giving an elasticity of about 0.15 by (10), compared with 0.45 in our illustrative model. Three conversations are far too few to draw conclusions from, and because the judges were models, Proposition 3 means part of the low scores could reflect a preference for length. We include the pilot because it is the only direct evidence we have, including the parts that do not favour us.

We are not reporting how many replies each average was computed over, how widely reply lengths varied, which prompts were used, or how Bloom was trained. Without the spread we cannot put an interval on either average.

09 Limitations and broader impact

We have two means, and two means do not describe a distribution. An average of 11 words could come from replies that are all close to 11, or from a mixture of three-word and forty-word replies, and Section 5 implies these would be very different companions.

We have not shown that the shorter replies are better. There is no quality measure on Bloom's own replies before and after, and the judges in the pilot may simply prefer longer text.

The model is deliberately simple. Propositions 1 and 2 assume value that saturates exponentially and a cost that is identical for every word. If each additional word costs more than the last, which could happen when a long reply crowds out the person's next turn, the asymmetry in Proposition 2 becomes weaker and may reverse. None of τ, B or κ has been fitted to data.

Words are a rough unit. Languages differ in how much a single word carries, not everyone writes in English, and all our counts are English words.

On broader impact, reply length affects who does the talking, and therefore whose thinking a conversation ends up containing. We think a companion that says less and leaves room for the person is the better default. One that only asks questions is a different matter, because it leaves the person doing all the work. Nothing in this note is a clinical claim.

10 What a brevity evaluation must report

Based on the above, an evaluation of brevity should report at least the following.

- The distribution of reply lengths (median, spread, and the share of replies over 30 words) in place of a single mean.
- Reply length plotted against message length, with the elasticity from (10).
- Floor share from (1) over the same exchanges.
- The mix of the moves in Table 1 before and after, so a fall in padding can be distinguished from a fall in substance.
- The ratio of reflections to questions, which exposes brevity achieved by asking.
- For each model judge, its preference for length estimated with (9), and scores adjusted accordingly.
- Ratings from people on at least a sample, because the bias in (8) belongs to model judges.
- A quality measure on the same replies. A length result published without one is hard to interpret, and this note is an example.

11 Conclusion

Bloom now replies in about eleven words where it used to use fifty-four. The person gets the floor back and spends about three seconds reading a reply where it used to take fourteen. For a companion we think that is a move in the right direction.

Whether eleven is the right number is a separate question that we cannot yet answer. The ideal length depends on how much there is to say, undershooting is costlier than overshooting, and the judge scores we would normally rely on lean toward longer text. For those reasons an average is a poor target and a poor thing to report. What remains to be done is to measure the distribution and the elasticity, to assess quality with length controlled for, and to ask people.

Disclosure. Bloom is a product of Mitwa AI, and this note was written by the lab that builds it. No user conversation is available to us, and none was used. Reproducibility. The results in Sections 4 to 7 are closed-form and proved in the text or in Appendix A, and Table 2 follows from the two averages. The replies behind the averages, the pilot conversations and the training method are not released with this note.

References

Ayers, J. W., Poliak, A., Dredze, M., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine, 183(6), 589–596. doi.org/10.1001/jamainternmed.2023.1838

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. Journal of Memory and Language, 109, 104047. doi.org/10.1016/j.jml.2019.104047

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), Syntax and Semantics 3: Speech Acts, 41–58. Academic Press.

Levinson, S. C. (2016). Turn-taking in human communication: Origins and implications for language processing. Trends in Cognitive Sciences, 20(1), 6–14. doi.org/10.1016/j.tics.2015.10.010

Levinson, S. C., and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. Frontiers in Psychology, 6, 731. doi.org/10.3389/fpsyg.2015.00731

Mitwa Lab (2026). Memory you can trust: The asymmetric cost of invented memories in AI companions. Technical note 01.

Moyers, T. B., Manuel, J. K., and Ernst, D. (2014). Motivational Interviewing Treatment Integrity coding manual 4.2.1. University of New Mexico. casaa.unm.edu/download/miti4_2.pdf

Niederhoffer, K. G., and Pennebaker, J. W. (2002). Linguistic style matching in social interaction. Journal of Language and Social Psychology, 21(4), 337–360. doi.org/10.1177/026192702237953

Saito, K., Wachi, A., Wataoka, K., and Akimoto, Y. (2023). Verbosity bias in preference labeling by large language models. arxiv.org/abs/2310.10076

A Appendix A: Proofs

Proposition 1.

Differentiating (3) gives $(B/\tau)\,e^{-\ell/\tau} - \kappa$, which is decreasing in ℓ, so (3) is concave with a single maximum where the derivative is zero: $e^{-\ell/\tau} = \kappa\tau/B$, that is, $\ell^* = \tau \ln(B/\kappa\tau)$, positive when $B > \kappa\tau$. Replacing B by $2B$ adds $\tau \ln 2$.

Proposition 2.

At ℓ^*, $B\,e^{-\ell^*/\tau} = \kappa\tau$. Hence the loss from a reply x words too long is $\kappa[x - \tau(1 - e^{-x/\tau})]$, and from a reply x words too short is $\kappa[\tau(e^{x/\tau} - 1) - x]$. Their difference is $\kappa[\tau(e^{x/\tau} - e^{-x/\tau}) - 2x] = 2\kappa[\tau \sinh(x/\tau) - x]$, which is positive because $\sinh y > y$ for $y > 0$, and increasing because $\cosh y > 1$.

Equation (5).

From (4), $d\ell^*/d\tau = \ln(B/\kappa\tau) - 1 = \ell^*/\tau - 1$. Multiplying by τ/ℓ^* gives $1 - \tau/\ell^*$, which is below one, and above zero whenever $\ell^* > \tau$.