Interaction of model size, context length and text genre on Turkish language models' prediction of human reading

An analysis of MultiplEYE Turkish data

Onur Keleş[1], Sezen Bektaş Yüksel[2], Servet Arda Kayhan[1], Feyza Budan[1], İris Beyza Deveci[1], Nazik Dinçtopal Deniz[2,3]

Boğaziçi University
[1]Department of Linguistics
[2]Department of Foreign Language Education
[3]M.A. Program in Cognitive Science

1st National Psycholinguistics Symposium
Middle East Technical University, Ankara
September 26, 2026

The MultiplEYE project

MultiplEYE is an EU-funded COST Action (CA21131, 2022–2026) that records eye movements while people read.

- **Whole texts** let us compare model predictions using all preceding text or a shorter window.
- Texts from **different genres**, each followed by **comprehension questions**.
- Texts in **39 languages** support comparisons across languages.

We collected the **Turkish data** to study language-model predictions and human reading.

Project and data collection: Jakobi et al. (2025). Text corpus: Kasperė et al. (2026).

The present study

	Research question
Context length	How much preceding text should a language model receive to explain reading times?
Model size	Does a larger language model explain reading times better?
Genre and comprehension	How does reading vary across genres, and how does reading behaviour relate to understanding?

Surprisal and reading

Surprisal expresses how unexpected a word is in its context.

$$S(w_i) = -\log P(w_i \mid w_1, \ldots, w_{i-1})$$

Less expected words are read more slowly. (Levy, 2008; Smith & Levy, 2013)

This holds across different languages, including Turkish. (Wilcox et al., 2023)

We estimate surprisal with a language model. The estimate depends on **how much preceding context** we give it and **which model** we use.

How much context do readers use?

Our first question concerns **how much preceding text to provide**.

Memory limits

Readers may retain an imperfect memory of earlier text. Surprisal estimates can incorporate this loss of context. (Futrell et al., 2020)

Restricted context in language models

Shorter context improved the fit to English and Japanese reading times, especially for larger models. (Kuribayashi et al., 2022)

We ask: does this advantage extend to Turkish, a verb-final, agglutinative language?

Model size and the fit to reading times

Alongside context length, we vary **model size**.

Better prediction, better fit

Earlier work linked better next-word prediction to a better fit to English gaze durations. (Goodkind & Bicknell, 2018)

Inverse scaling

Later studies found a poorer fit for larger models in English and across ten languages, including Turkish. (Oh & Schuler, 2023; de Varda & Marelli, 2023)

Rare-word probabilities may help explain this pattern. (Oh, Yue & Schuler, 2024)

We ask: does the smaller-model advantage hold within a Turkish model family, including for later reading measures?

Text, re-reading and comprehension

Our third question concerns **readers' behaviour and understanding**. MultiplEYE lets us relate both to text genre.

Re-reading and text difficulty

More difficult texts lead to longer fixations and more returns to earlier text. (Rayner et al., 2006)

Preventing re-reading reduces comprehension. (Schotter et al., 2014)

Our questions:

- Does re-reading vary across genres?
- Within each genre, are re-reading and surprisal sensitivity associated with better comprehension?

Introduction

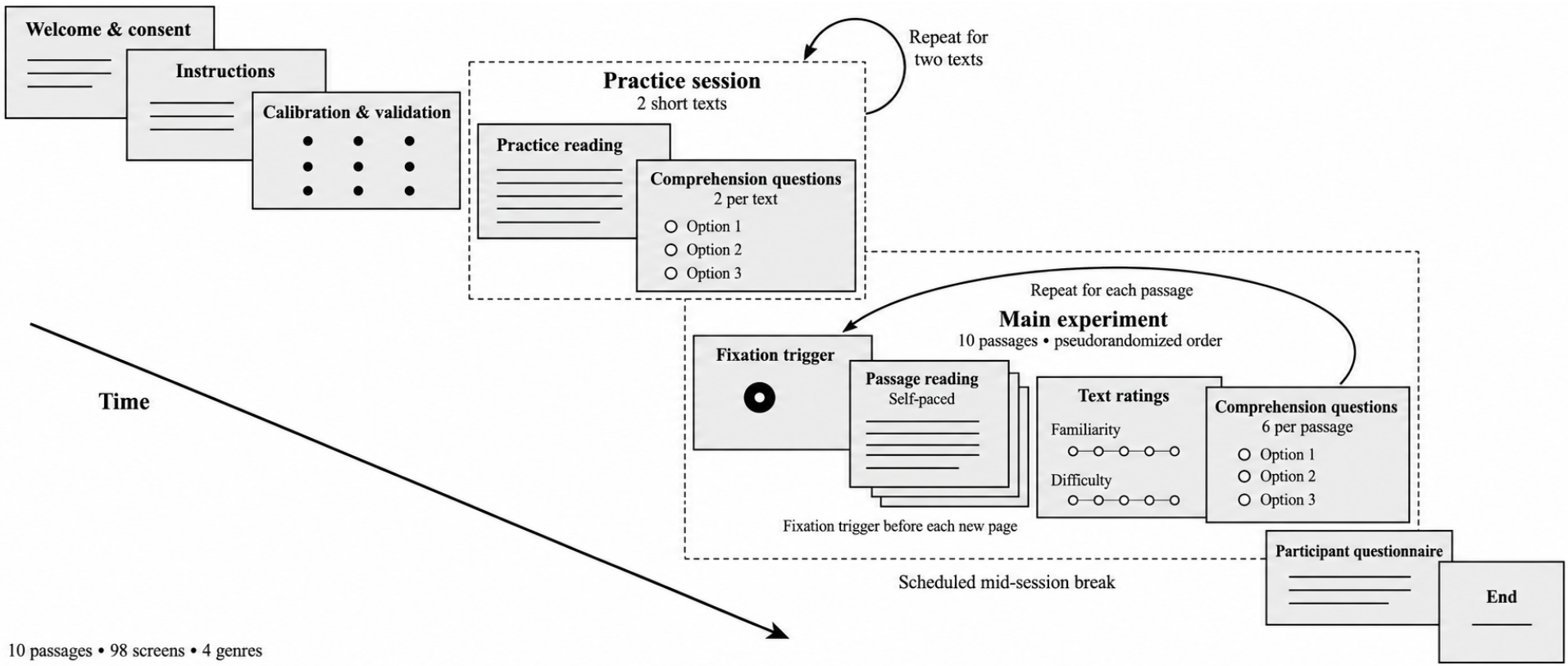
Methods

Results

Discussion

Participants and procedure

57 native Turkish readers; 10 passages (4,702 words) from four genres: Literary, Argumentative, Institutional, Popular Science.



Procedure follows the MultiplEYE data collection guidelines (Jakobi et al. (2025)).

Language models

Kanarya-750m and Kanarya-2b
Two sizes of the same Turkish language-model family: 750 million and 2 billion parameters.

In our screening of seven Turkish models, Kanarya had the lowest average number of subword tokens: **1.35 per word**.

Comparing two sizes within one family lets us examine the relationship between model size and fit to reading times.

Tokenizer fertility: Rust et al. (2021). Models: Safaya et al. (2022) and the Kanarya model documentation.

Context conditions

We vary the context given to the **language model**. Readers see the same text in every comparison.

Dün akşam Ayşe kütüphanede sınava hazırlanan arkadaşına yeni kitabını ***verdi****.*

Target: **verdi** (gave)

Condition	Preceding words available to the model
Full	Dün akşam Ayşe kütüphanede sınava hazırlanan arkadaşına yeni kitabını
7-gram	kütüphanede sınava hazırlanan arkadaşına yeni kitabını
2-gram	kitabını

Windows tested: 20, 10, 7, 5, 3 and 2-gram (*n* includes the target word).

Constructed example. Context definition: Kuribayashi et al. (2022), Section 3.2.

Eye-movement measures

	Measure	Definition
FFD	First fixation duration	Duration of the first fixation on a word
GD	**Gaze duration**	Sum of first-pass fixations before the eyes leave the word
TD	**Total duration**	Sum of all fixations on the word, including later returns
RRD	**Re-reading duration**	Time spent on the word after the first pass
TRCin	Incoming regressions	Number of regressions into the word
PRO	Regression-out probability	Probability of a regression out of the word during the first pass

We focus on **GD** (first pass), **TD** (all visits), **RRD** (later visits), and the proportion of words re-read.

Gaze-duration definition: Goodkind & Bicknell (2018).

Statistical model

For each reading-time measure, language model and context condition, we fit a linear mixed-effects model:

$$\log \mathrm{RT} = B_w + \underbrace{\gamma_0 S_w + \gamma_1 S_{w-1} + \gamma_2 S_{w-2}}_{\text{current and preceding words' surprisal}} + u_{\text{reader}} + v_{\text{passage}} + \varepsilon$$

B_w: intercept, word length, frequency and their interactions for the current and previous word.

Reader and passage intercepts account for differences in average reading time. Adding passage intercepts improves fit:

Likelihood-ratio test	Gaze	Total	Re-reading
$\chi^2(1)$	1596	4002	419
p	$< .001$	$< .001$	$< .001$

lme4, maximum likelihood; Kanarya-750m, full context (Kanarya-2b: same pattern). S_{w-1}, S_{w-2}: spillover. Frequency: the model's context-free word probability.

Methods

Measuring predictive power

Each statistical model gives two quantities:

- **The surprisal effect**: the change in log reading time per extra nat (γ_0).
- **Psychometric predictive power (PPP)**: the gain in model fit from adding the three surprisal terms.

$$\text{PPP} = 1000 \, \frac{\log L_{\text{with surprisal}} - \log L_{\text{without surprisal}}}{N}$$

Higher PPP means adding surprisal improves the fit more. We compare this gain across context conditions and language models.

Goodkind & Bicknell (2018) and Kuribayashi et al. (2022). L: likelihood; N: observations. PPP measures in-sample fit, not held-out prediction or variance explained.

Surprisal and reading times

Less predictable words are read more slowly.

Per extra nat of surprisal	Change	ms	t, p
Gaze duration	$+4.3\%$	$+9.5$	$40.3, < .001$
Total duration	$+7.2\%$	$+19.6$	$53.6, < .001$

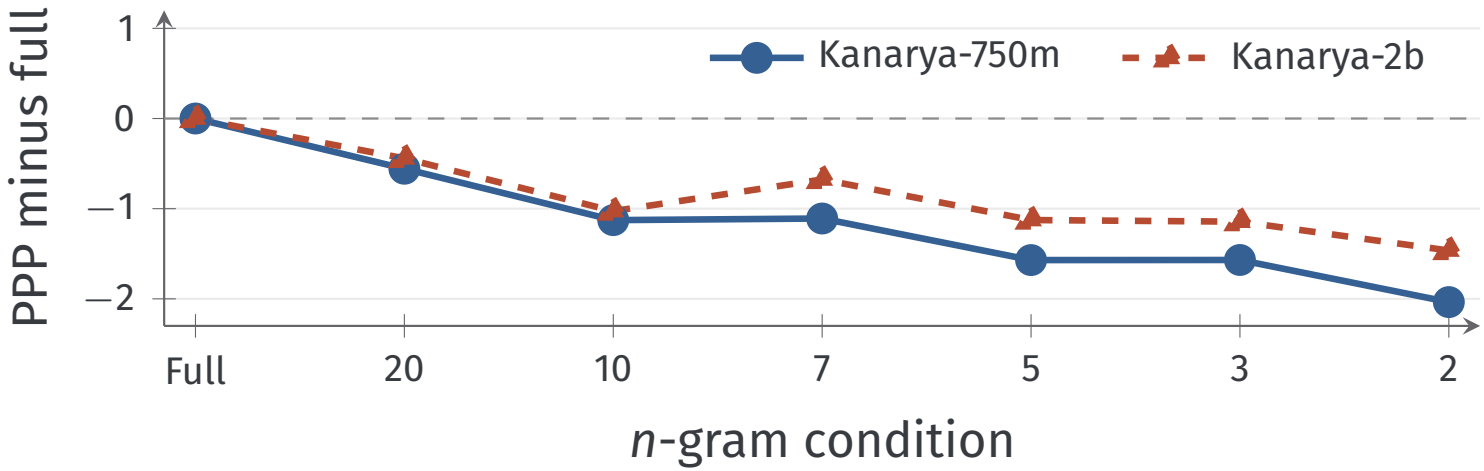
A tenfold decrease in word probability corresponds to about **22 ms longer** on the first pass and **47 ms longer** in total, at the reference durations below.

This is consistent with previous evidence from eleven languages, including Turkish. Next, we ask which context window gives the best fit.

Kanarya-750m, full context. Converted at typical durations: 219 ms (gaze), 271 ms (total). Wilcox et al. (2023).

Context length

With reader and passage intercepts. Gaze-duration PPP relative to full context. Above zero favours restricted context.



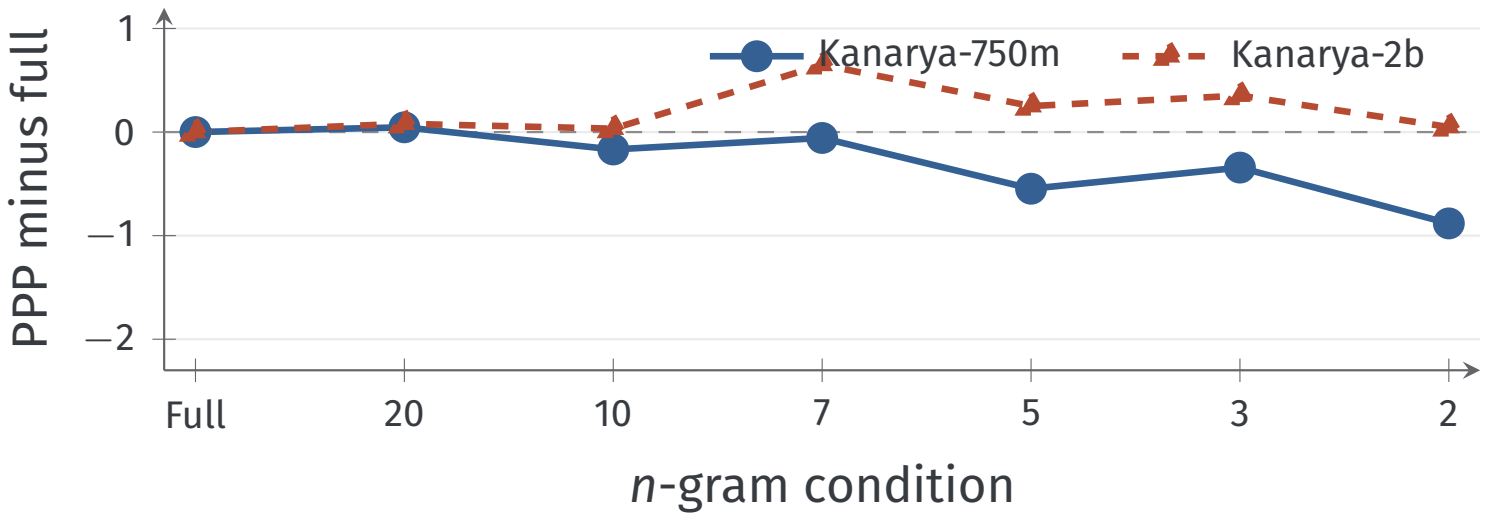
7-gram minus full	Kanarya-2b	Kanarya-750m
Gaze duration	-0.68, $p = .007$	-1.11, $p < .001$
Total duration	-1.10, $p < .001$	-2.09, $p < .001$
Re-reading duration	-0.26, $p = .318$	-0.92, $p = .007$

No restricted window has higher PPP than full context.

Holm-corrected paired tests.

Context length

Now omit passage intercepts. Reader intercepts remain. Above zero favours restricted context.



Kanarya-2b now fits better in the 7-gram condition: gaze ($p = .033$), total ($p < .001$) and re-reading duration ($p = .044$). Which passage differences account for this reversal?

Holm-corrected paired tests. Kanarya-750m: no significant 7-gram advantage.

Context length

Preamble of the Universal Declaration of Human Rights, with Kanarya-2b surprisal (nats):

İnsanlık 7.4 ailesinin 0.6 bütün 0.8 üyelerinde 0.7 bulunan 0.1 haysiyetin 0.6 ve 0.0 bunların 0.0 eşit 0.1 ve 0.0 devir 0.0 kabul 0.2 etmez 0.0 haklarının 0.0 tanınması 0.0 hususunun, 0.0 hürriyetin, 3.6 adaletin 0.2 ve 0.0 dünya 0.0 barışının 0.1 temeli 0.1 olmasına, 0.7 …

Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world, …

Mean, full context	Human Rights	Other nine passages
Surprisal, Kanarya-2b / 750m	**0.37** / **1.28**	3.9 to 5.9
Readers' gaze duration	**250 ms** (slowest)	216 to 244 ms

The models find this passage highly predictable, yet readers read it most slowly. This suggests **possible training-data contamination**.

Surprisal versus the other passages, word level: Mann–Whitney $p < .001$ for both models.

Context length

Readers-only model, refitted without one passage at a time

7-gram minus full-context PPP, gaze duration

Passage excluded	Kanarya-2b	Kanarya-750m
Any one of the other nine	$+0.50$ to $+1.12$	-0.22 to $+0.45$
Human Rights	$-\mathbf{1.21}$, $p < .001$	$-\mathbf{1.97}$, $p < .001$

Within the Human Rights passage (Kanarya-2b, gaze-duration PPP): full context 0.15, 2-gram 6.43 ($p < .001$).

Removing this passage reverses the 7-gram advantage. One explanation is that a short window limits the use of memorised text. Memorisation remains an inference from model behaviour.

Model size

Full context for both models

Both use the same external word-frequency control, so the PPP comparison has a common baseline.

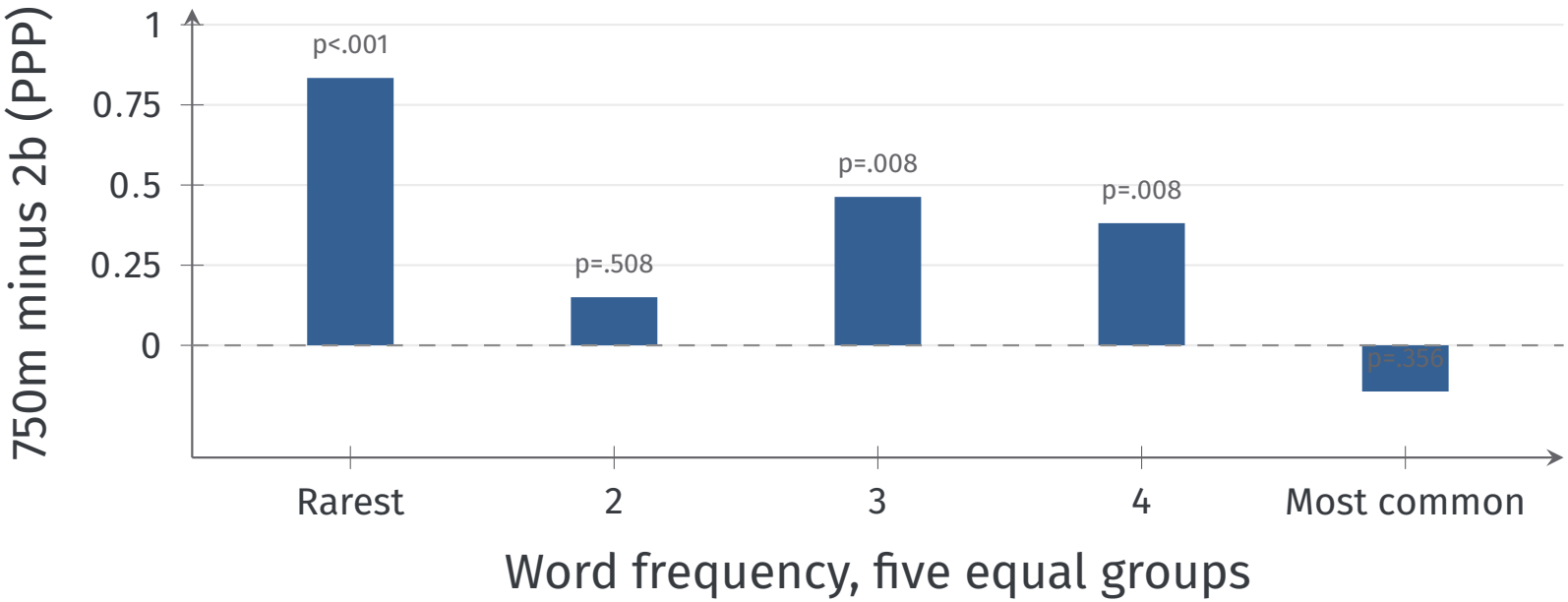
PPP	Kanarya-750m	Kanarya-2b	p
Gaze duration	**4.47**	4.03	$< .001$
Total duration	**8.30**	7.72	$< .001$
Re-reading duration	**2.94**	2.56	$< .001$

The smaller model fits better on all three measures.

The smaller-model advantage extends to this Turkish model family.

Paired tests. External frequency measure (wordfreq) so that both models share one baseline.

Model size



Above zero favours the smaller model. Labels give *p*. Its advantage is **largest for the rarest words**. We detect no advantage for the most common words, consistent with the proposed role of word frequency.

Gaze duration. Frequency account: Oh, Yue & Schuler (2024).

Genre and comprehension

	Literary	Popular science	Argumentative	Institutional
Words re-read	26%	22%	19%	**30**%
Reading time per word	247 ms	223 ms	198 ms	**269** ms
Share of time after first pass	35%	31%	29%	**39**%
Comprehension accuracy	63%	67%	70%	69%

Re-reading varies by genre ($\chi^2(3) = 8.9$, $p = .031$), controlling for word length and frequency, with reader and passage intercepts. In these texts, it is most frequent in Institutional and least frequent in Argumentative passages.

Genre and comprehension

Within each genre, we relate comprehension accuracy to:

- **Re-reading**: the proportion of words each reader re-read.
- **Surprisal sensitivity**: each reader's correlation between word surprisal and total reading time.

Genre	Re-reading	Surprisal sensitivity
Literary	$\mathbf{r = .43}$, $p = .001$ (.003)	$r = .27$, $p = .045$ (.134)
Popular Science	$\mathbf{r = .38}$, $p = .004$ (.008)	$\mathbf{r = .44}$, $p < .001$ (.002)
Argumentative	$r = .25$, $p = .057$ (.057)	$r = .10$, $p = .469$ (.469)
Institutional	$\mathbf{r = .44}$, $p < .001$ (.002)	$r = .16$, $p = .225$ (.450)

Re-reading correlates with comprehension in three genres. For surprisal sensitivity, only Popular Science meets the threshold after correction. These tests do not compare genres' correlations.

57 readers. p unadjusted; in parentheses, Holm-adjusted. Bold: significant after adjustment.

Introduction

Methods

Results

Discussion

Context length: the Human Rights passage

Full context fits best

Once passage intercepts are included, no shorter window improves the fit.

The short-window advantage comes from one passage

Kanarya-2b predicts the preamble almost perfectly, yet readers read it most slowly. A short window lowers the model's predictability of this passage (PPP from 0.15 to 6.43).

Why memorisation?

Larger models memorise more, and longer context helps them recall memorised text (Carlini et al., 2023). Only Kanarya-2b shows the advantage.

Context length: relation to earlier work

Not replicated in Turkish

Earlier work found the advantage in English and Japanese, with article intercepts. Japanese is also verb-final, so word order does not readily explain the difference.

Multilingual texts may be more exposed

English reading corpora overlap little with training data (Oh, Zhu & Schuler, 2025). Multilingual corpora include widely copied texts, such as the Human Rights declaration.

Lossy memory is not ruled out

A gradual recency bias improves fit in English (Clark et al., 2025); cutting context to *n* words may be too blunt a test.

Kuribayashi et al. (2022); Futrell et al. (2020).

Why Kanarya?

We also computed surprisal with five other Turkish models. Of seven Turkish models, only Kanarya ranks the Human Rights passage as the most predictable.

Model	Training	Rank
Kanarya-750m, 2b	Turkish web text (OSCAR, mC4); no deduplication reported	**1st, 1st**
COSMOS GPT-2 (3 sizes)	Mainly deduplicated CulturaX (mC4, OSCAR)	7th, 4th, 4th
Turkish-Llama-8b	Llama 3 (deduplicated), then Turkish	6th
Turkish-Gemma-9b	Gemma-based, then Turkish	6th

The declaration is copied on many Turkish websites. Repeated text is memorised far more often (Kandpal et al., 2022), and deduplication removes these copies.

Rank: lowest mean surprisal of the ten passages = 1st. Training details from model cards and papers; exact training data are not public.

Model size

Inverse scaling within one Turkish model family

The smaller model fits better on all three measures, including total and re-reading duration. This holds without the Human Rights passage.

Readers' expectations come from their experience

In surprisal theory, the relevant probabilities are the reader's, learned from exposure (Levy, 2008). Larger models predict rare words and names better than readers do, which fits our rare-word result. Models closer to readers' experience fit better (Škrjanec & Demberg, 2025).

Oh & Schuler (2023); de Varda & Marelli (2023); Oh, Yue & Schuler (2024); Oh, Zhu & Schuler (2025).

Genre and comprehension

Register affects reading

Institutional texts were re-read most. Legal and official registers tend to contain dense noun phrases and nominalised clauses (Biber & Conrad, 2009).

Re-reading goes with comprehension

In three genres, readers who re-read more scored higher. Returning to earlier text may help readers build a coherent representation of the text (Kintsch, 1998); preventing it lowers comprehension.

Surprisal sensitivity

It reaches significance for Popular Science. Readers know less about science topics, so comprehension depends more on the text itself (Ozuru et al., 2009). Unexpected words carry the most new information, so slowing down on them may help most in these texts.

Rayner et al. (2006); Schotter et al. (2014).

Conclusions

1. Full context gives the best fit to Turkish reading times. The short-window advantage depends on one likely memorised passage.
2. A smaller model fits reading times better, especially for rare words.
3. Re-reading varies by genre and goes with better comprehension.

Take-home: control for passage differences and check for memorised texts, especially in larger models trained on web data without near-duplicate removal.

Onur Keleş onur.keles1bogazici.edu.tr

Appendix: Tokenizer screening

Fertility: mean subwords per word.

Checkpoint or group	Fertility
Kanarya-750m and Kanarya-2b	1.35
Turkish GPT-2 (base)	1.88
Turkish-Gemma-9b-T1	2.34
Turkish-Llama-8b-v0.1	2.54
tr-Qwen2.5-0.5B-SFT-v1	2.74

Lower fertility means less subword splitting.

Tokenizer screening. Rounded to two decimals. Definition: Rust et al. (2021).

References: surprisal and reading

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.
https://doi.org/10.1016/j.cognition.2007.05.006

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319.
https://doi.org/10.1016/j.cognition.2013.02.013

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the ACL, 11*.
https://doi.org/10.1162/tacl_a_00612

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *CMCL 2018*, 10–18.
https://doi.org/10.18653/v1/W18-0102

References: context, text and comprehension

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814.
https://doi.org/10.1111/cogs.12814

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022). Context limitations make neural language models more human-like. *EMNLP 2022*, 10421–10436.
https://doi.org/10.18653/v1/2022.emnlp-main.712

Kuribayashi, T., Oseki, Y., Ben Taieb, S., Inui, K., & Baldwin, T. (2025). Large language models are human-like internally. *Transactions of the ACL*.
https://doi.org/10.1162/TACL.a.58

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241–255.
https://doi.org/10.1207/s1532799xssr1003_3

Schotter, E. R., Tran, R., & Rayner, K. (2014). Don't believe what you read (only once): Comprehension is supported by regressions during reading. *Psychological Science*, *25*(6), 1218–1226.
https://doi.org/10.1177/0956797614531148

References: model size and contamination

de Varda, A. G., & Marelli, M. (2023). Scaling in cognitive modelling: A multilingual approach to human reading times. *ACL 2023, Short Papers*.
https://aclanthology.org/2023.acl-short.14/

Oh, B.-D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the ACL, 11*.
https://doi.org/10.1162/tacl_a_00548

Oh, B.-D., Yue, S., & Schuler, W. (2024). Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. *EACL 2024*.
https://aclanthology.org/2024.eacl-long.162/

Oh, B.-D., Zhu, H., & Schuler, W. (2025). The inverse scaling effect of pre-trained language model surprisal is not due to data leakage. *Findings of ACL 2025*.
https://aclanthology.org/2025.findings-acl.91/

References: corpus and models

Jakobi, D. N., et al. (2025). MultiplEYE: Creating a multilingual eye-tracking-while-reading corpus. *ETRA '25*, Article 111.
https://doi.org/10.1145/3715669.3726843

Kasperė, R., Bondar, A., Nisioi, S., et al. (2026). The MultiplEYE Text Corpus: Towards a diverse and ever-expanding multilingual text corpus. *LREC 2026*.
https://aclanthology.org/2026.lrec-1.534/

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How good is your tokenizer? On the monolingual performance of multilingual language models. *ACL-IJCNLP 2021*, 3118–3135.
https://doi.org/10.18653/v1/2021.acl-long.243

Safaya, A., Kurtuluş, E., Goktogan, A., & Yuret, D. (2022). Mukayese: Turkish NLP strikes back. *Findings of ACL 2022*, 846–863.
https://doi.org/10.18653/v1/2022.findings-acl.69

Model and conference documentation

Kanarya model documentation. Model cards accompanying the two checkpoints. Accessed September 21, 2026.
https://huggingface.co/asafaya/kanarya-750m

https://huggingface.co/asafaya/kanarya-2b

Conference information. 1. Ulusal Psikodilbilim Sempozyumu, September 25–26, 2026.

https://sites.google.com/view/odtupds/ana-sayfa