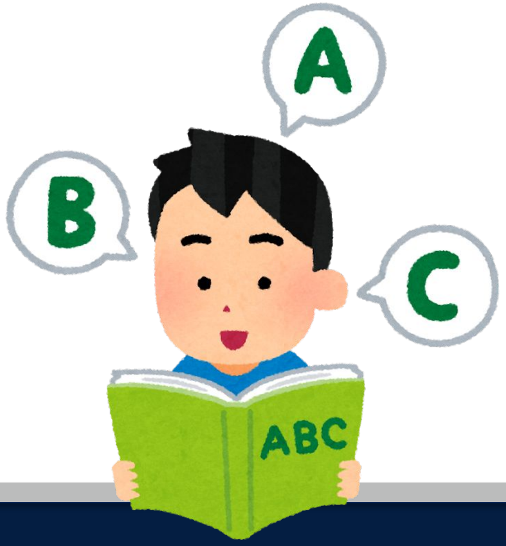
GEORGETOWN UNIVERSITY

Modeling Non-Native Sentence Processing with L2 Language Models

Tatsuya Aoyama & Nathan Schneider
Georgetown University

High Level Overview

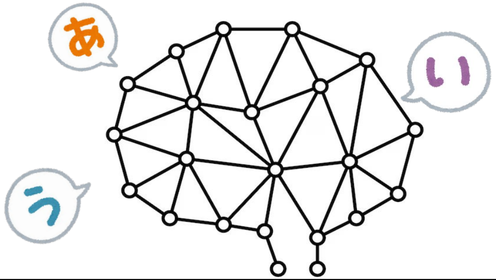
- For non-native speakers of English, first language (L1) affects many aspects of second language (L2) performance… including morphosyntactic knowledge (Murakami & Alexopoulou, 2016) and **sentence processing** (Clahsen & Felser, 2006)



High Level Overview

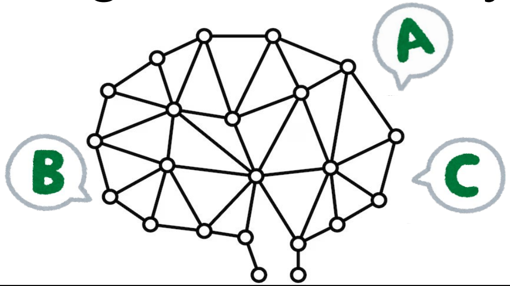
Different L1s, same L2 (English)
Second language language models (**L2LMs**)

- Do LMs with different “L1s” also read English differently?



Pierre Vinken, 61 years old, will join…

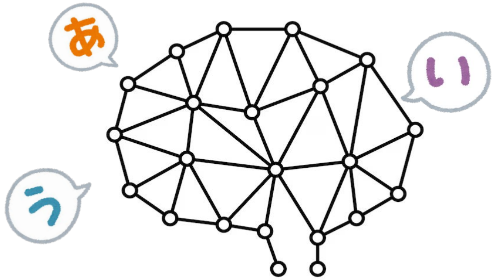
≠



Pierre Vinken, 61 years old, will join…

?

- Does their sentence processing match that of human with the same L1?



Pierre Vinken, 61 years old, will join…

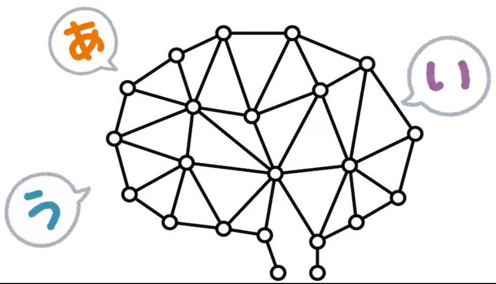
=

Pierre Vinken, 61 years old, will join…

?

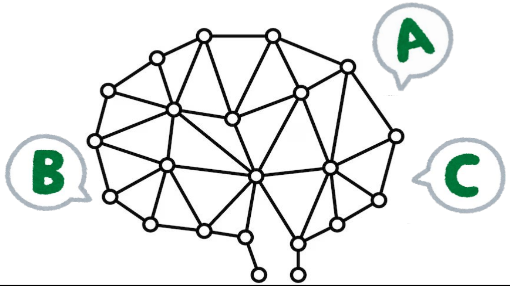
High Level Overview

- They do read differently!



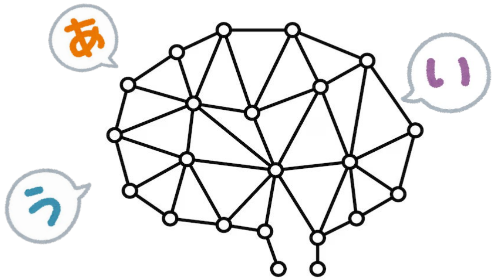
Pierre Vinken, 61 years old, will join…

≠



Pierre Vinken, 61 years old, will join…

- But not exactly like humans with the same L1!



Pierre Vinken, 61 years old, will join…

≠

Pierre Vinken, 61 years old, will join…

Related Work

- LMs as models of human language acquisition
  - BabyBERTa (Huebner et al., 2021)
  - BabyLM challenge (Warstadt et al., 2023; Choshen et al., 2024)
  - SLA (Yadavalli et al., 2023, Oba et al., 2023)
    - **Test for Inductive Bias via Language Model Transfer** (TILT; Papadimitriou & Jurafsky, 2020)

L2LMs with the same L2 (English)

- LMs and sentence processing
  - Surprisal theory (Hale 2001, Levy 2008)
  - LMs' "psychometric predictive power" (PPP)
  - PPP *positively* correlates with LM quality (*quality-power hypothesis*; Wilcox et al., 2020, Wilcox et al., 2023) until certain point in pretraining (Oh & Schuler, 2023)

Data & Pretraining

- L1s: English, Spanish, Portuguese, Arabic, Chinese, Japanese
- L2: English
- Model: GPT-2

Data & Pretraining

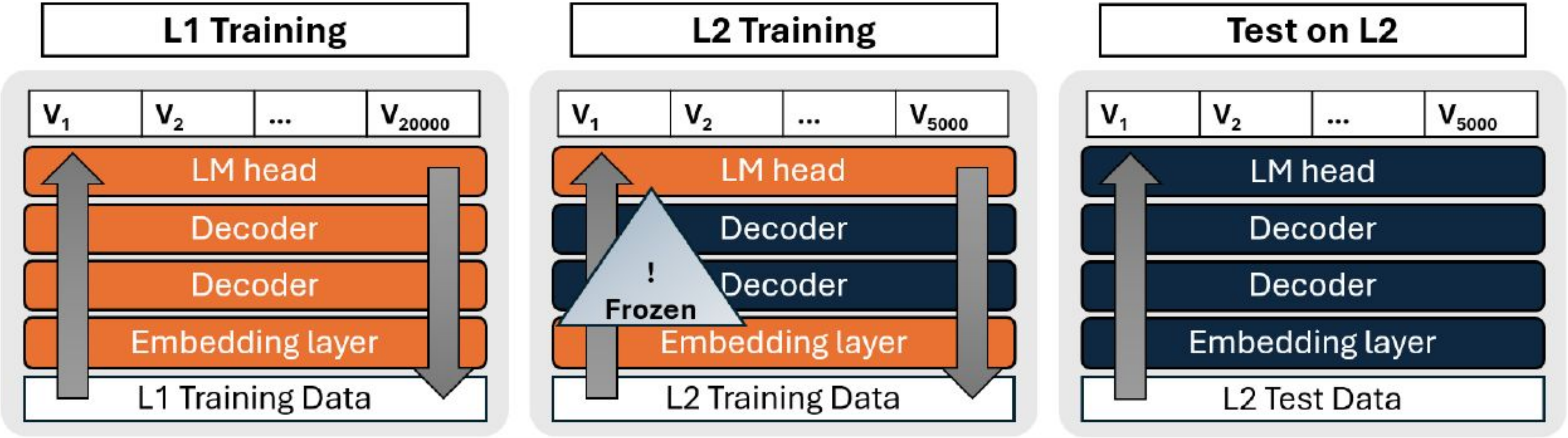


Figure 1: Training setup for L2LMs. The model is first pretrained on a given L1. We then freeze all the layers *except for* the embedding and output layers, and then continue pretraining on the L2 (English).

Data & Pretraining

- First language
  - CC100 (sampled 100M tokens)
- Second language
  - Simple English Wikipedia (sampled 30M tokens)
- Reading time data
  - CELER (Berzak et al., 2020): eye-tracking data from participants with 6 L1 backgrounds

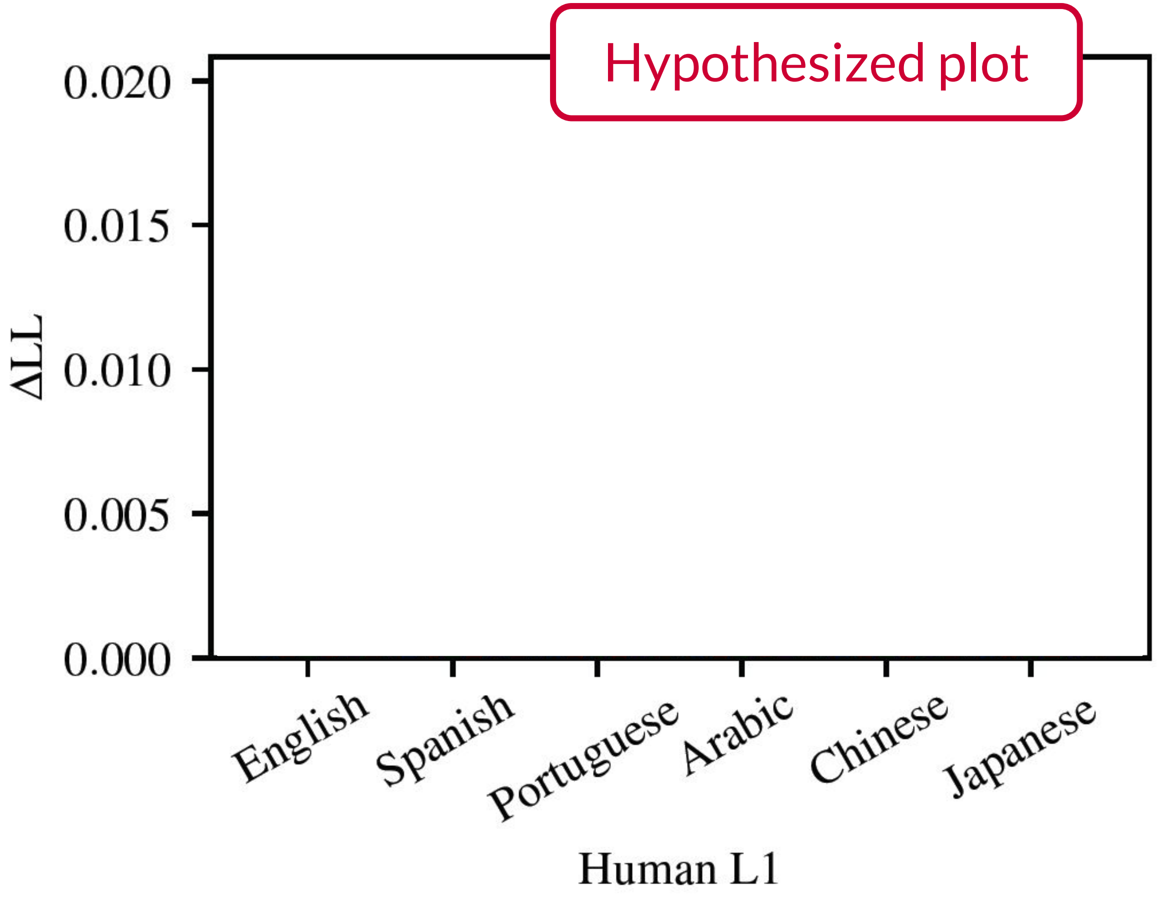
Evaluation

L1 effects were observed!
Check our paper for these results!

RQ1
1. L2 perplexity (PPL)
2. L2 grammatical knowledge (BLiMP; Warstadt et al., 2020)

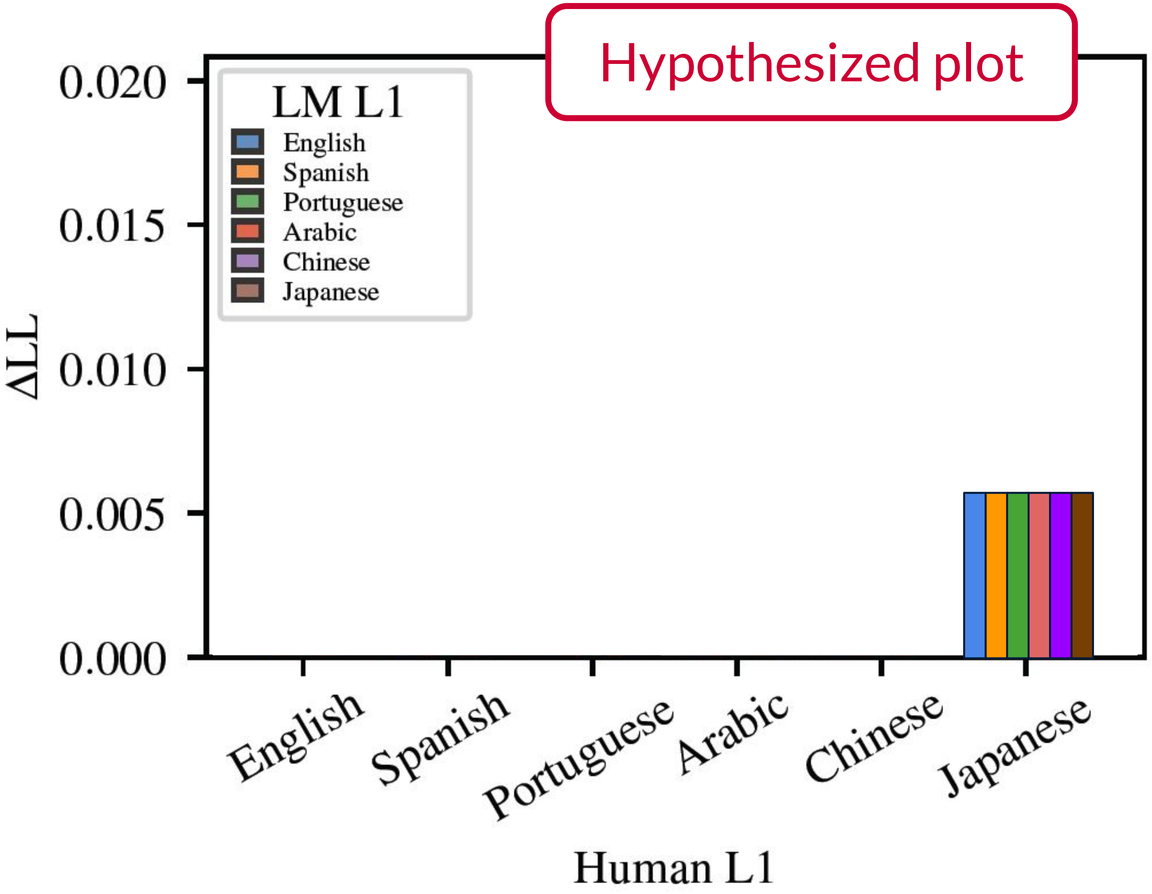
RQ2
3. **L2 sentence processing**
   - Compare 2 linear regression models (baseline model vs baseline+surprisal model)
   - Surprisal: $$S_{w_i} = -\log P(w_i \mid \boldsymbol{w}_{<i}). \quad (1)$$
   - ΔLL: $$\Delta LL = LL_{\phi_{bl+S}} - LL_{\phi_{bl}}, \quad (2)$$

RQ2



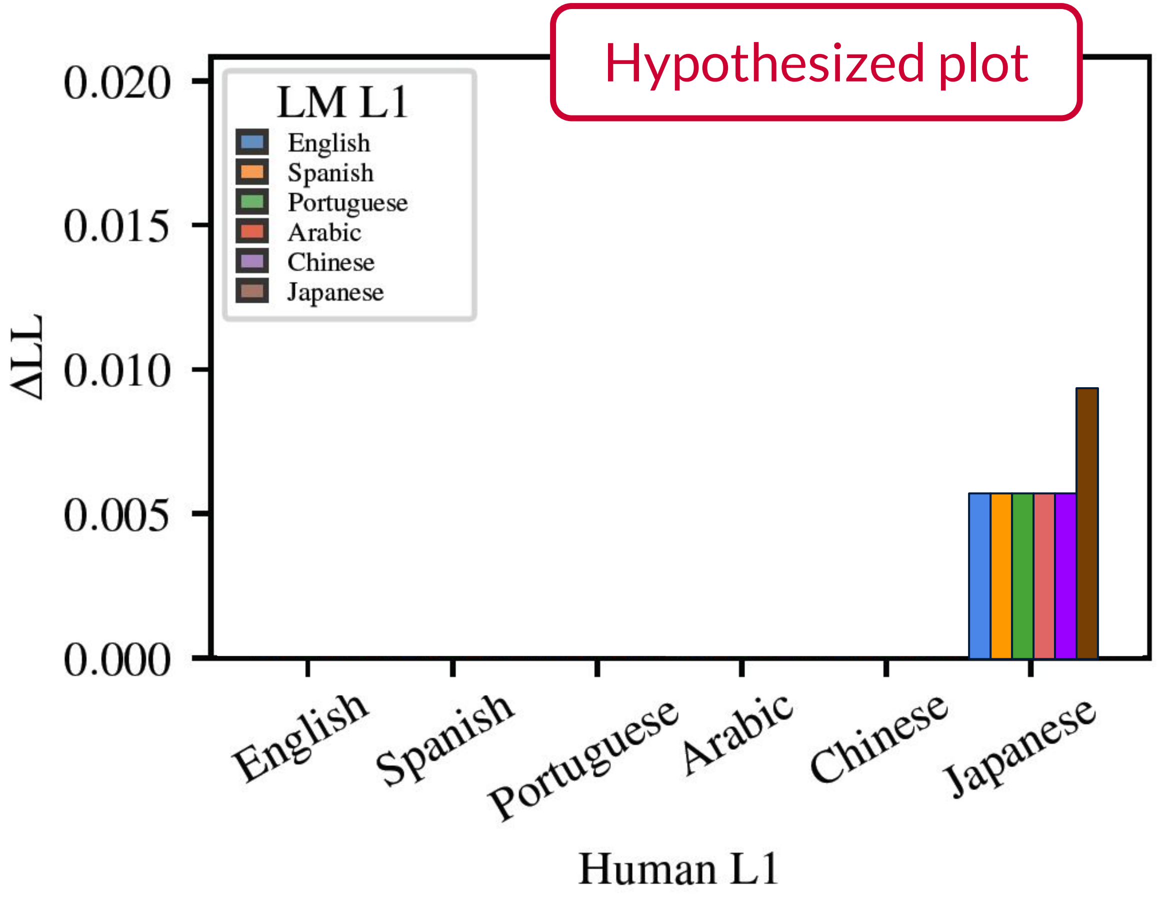
- X axis: Human L1
- Bar color: LM L1
- Y axis: ΔLL (LM-human alignment)

RQ2



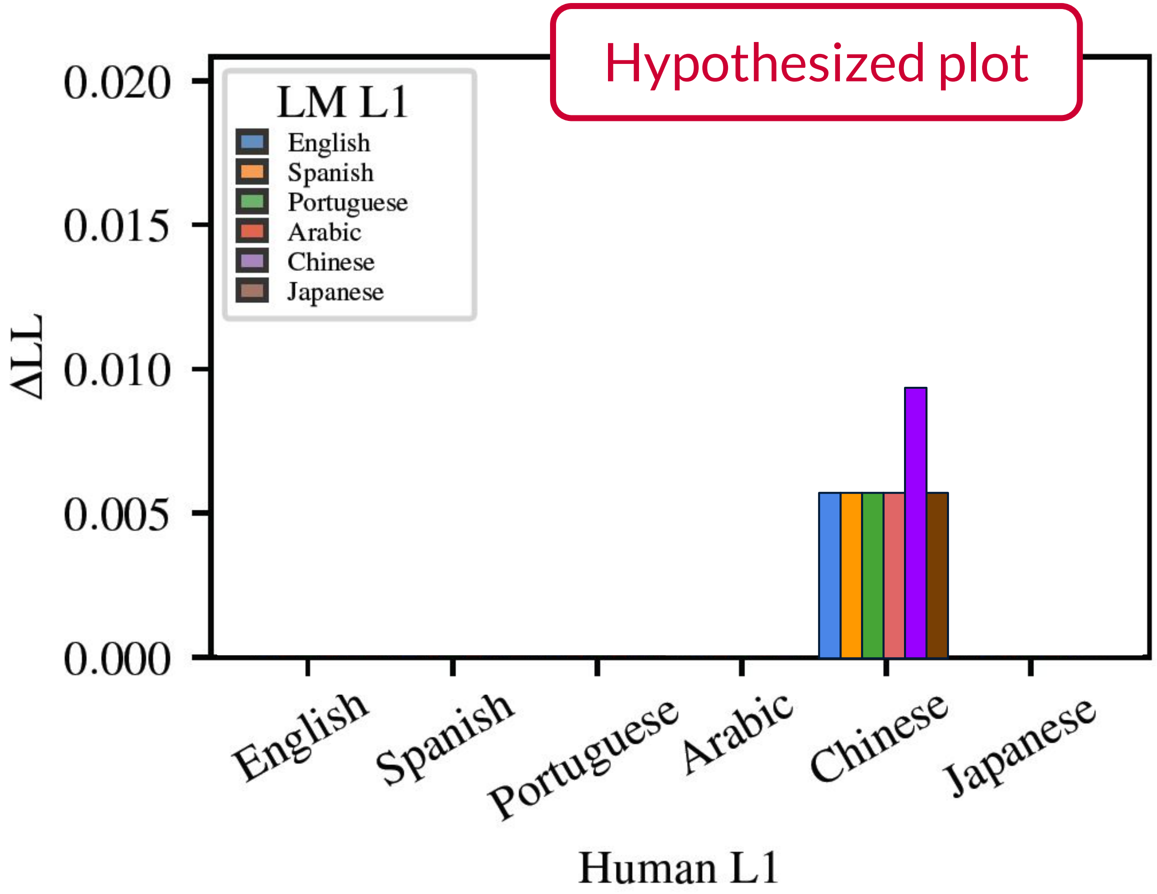
- If LM's L1 does not matter…

RQ2



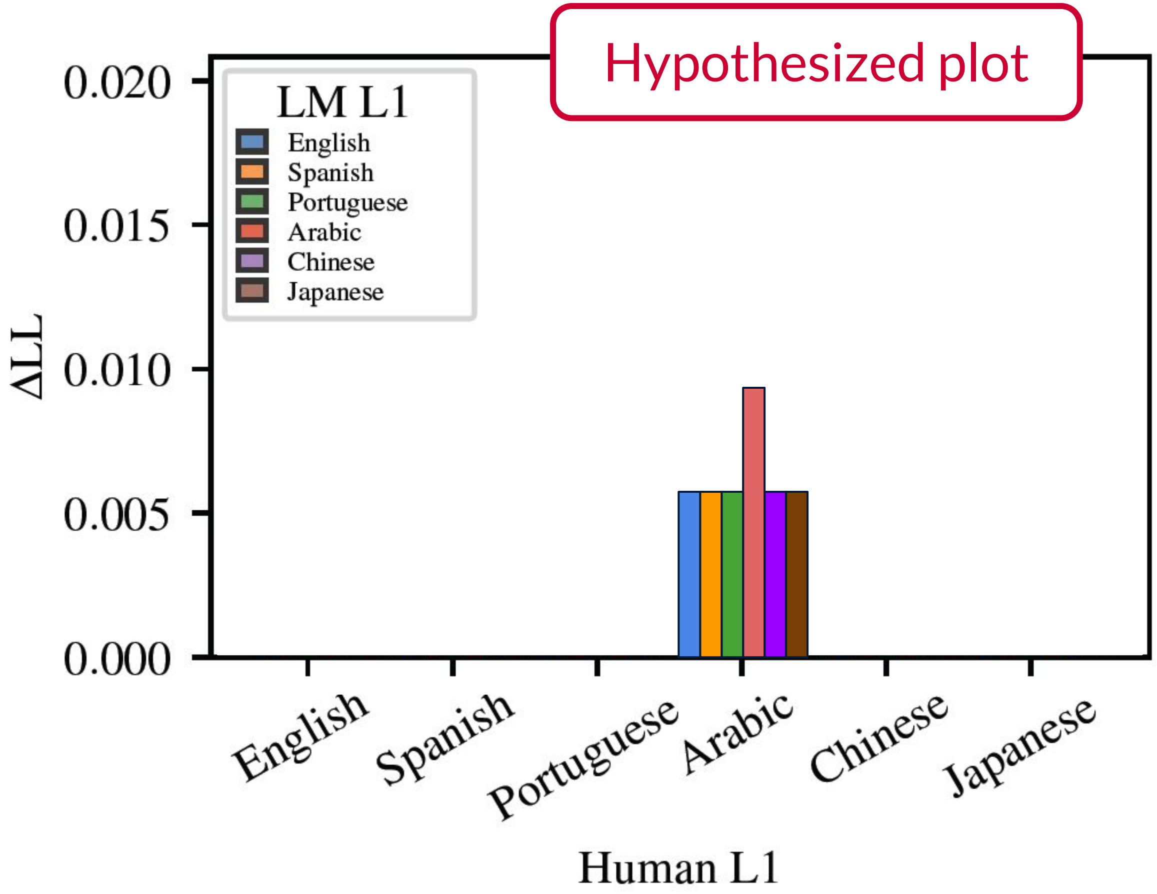
- If it does matter, and **matching the L1** result in the highest ΔLL (this is our hypothesis)

RQ2



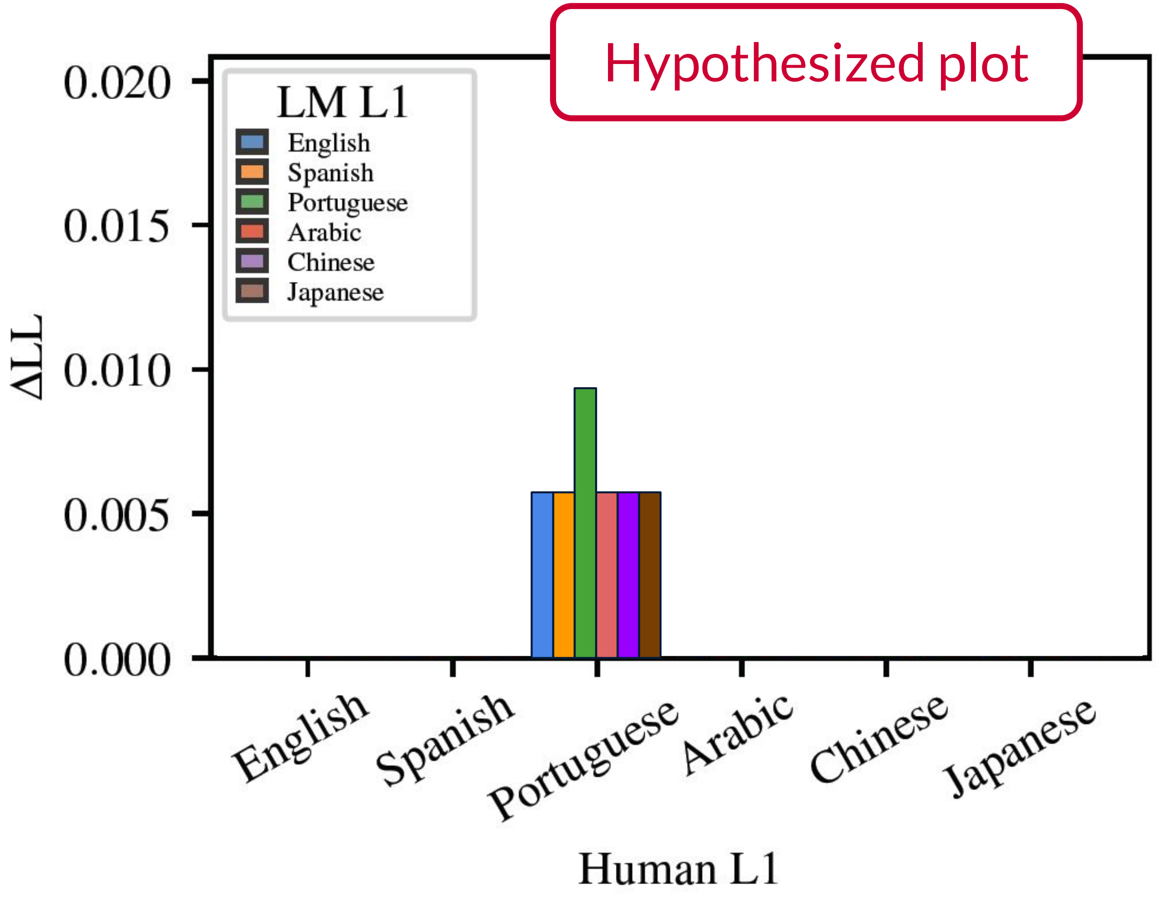
- This should apply to other languages as well

RQ2



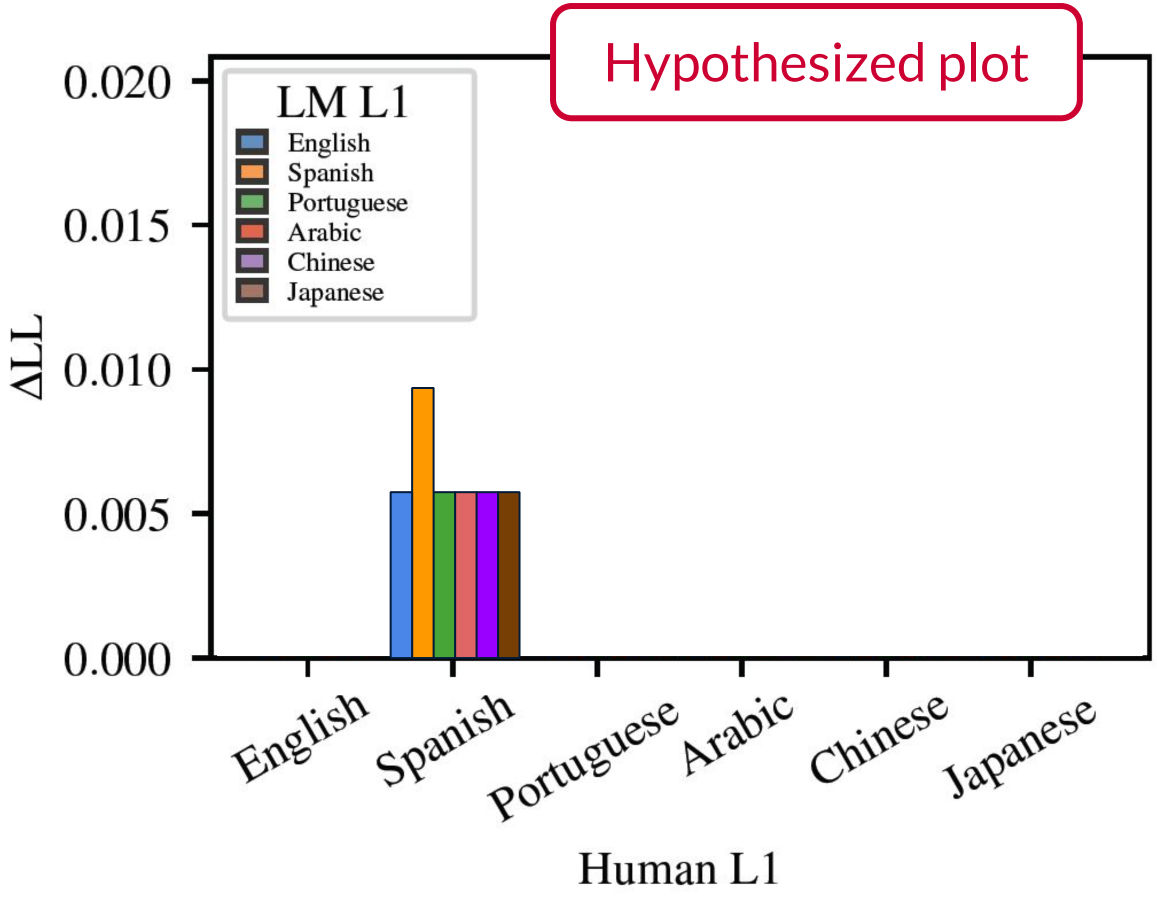
- This should apply to other languages as well

RQ2



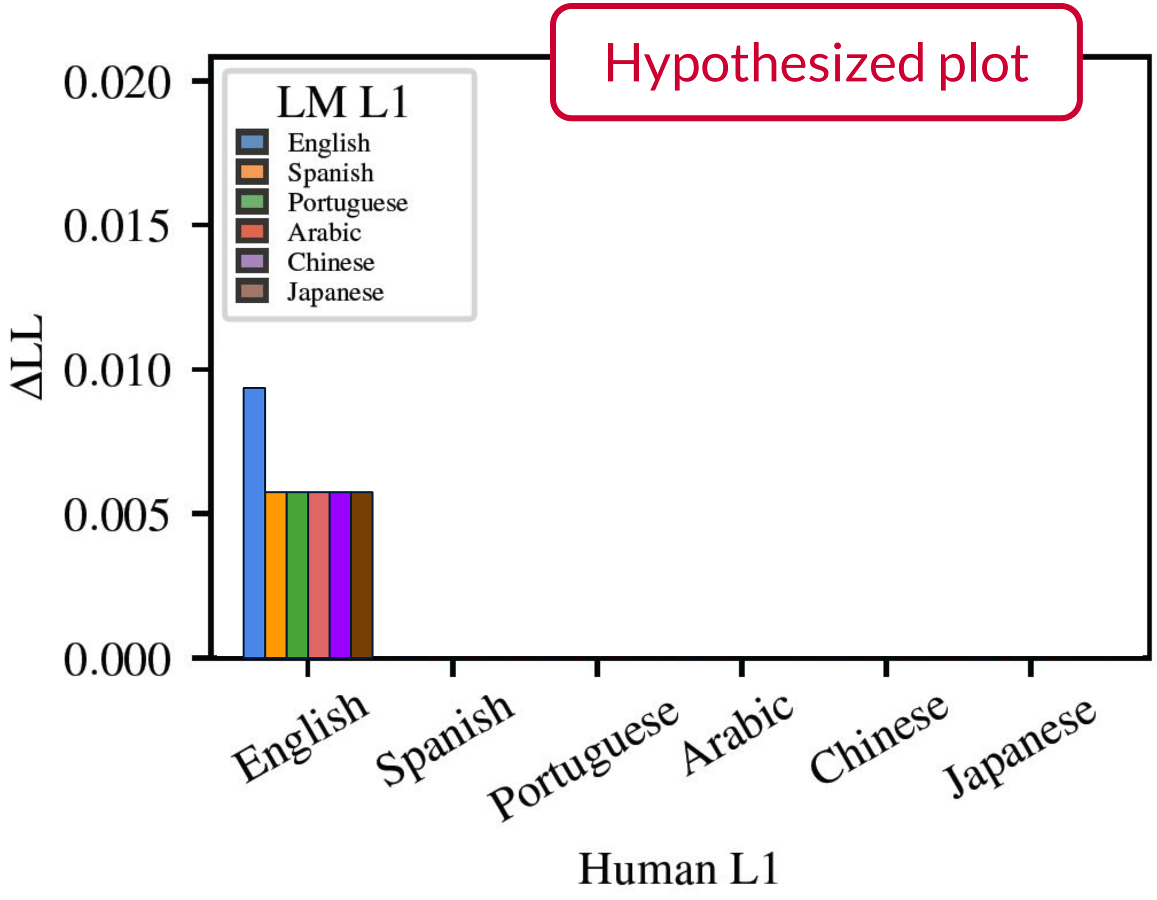
- This should apply to other languages as well

RQ2



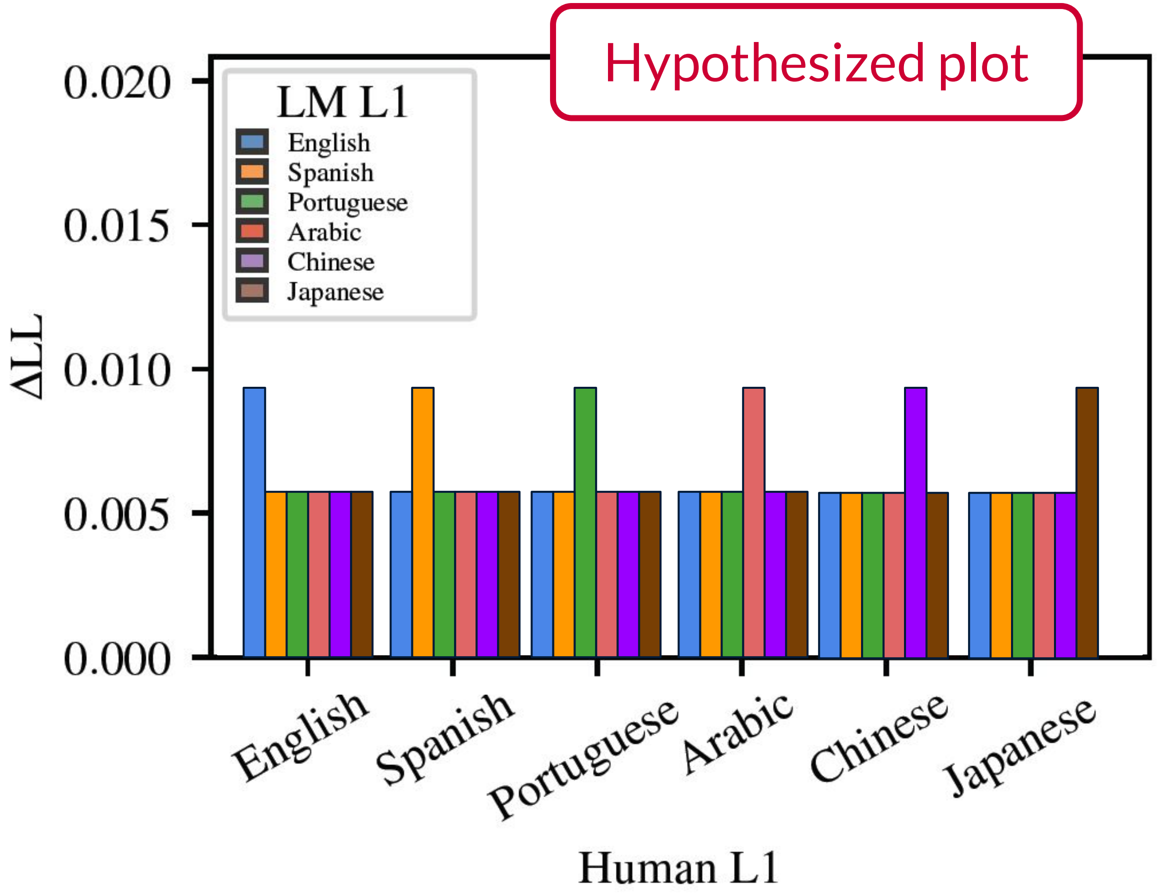
- This should apply to other languages as well

RQ2



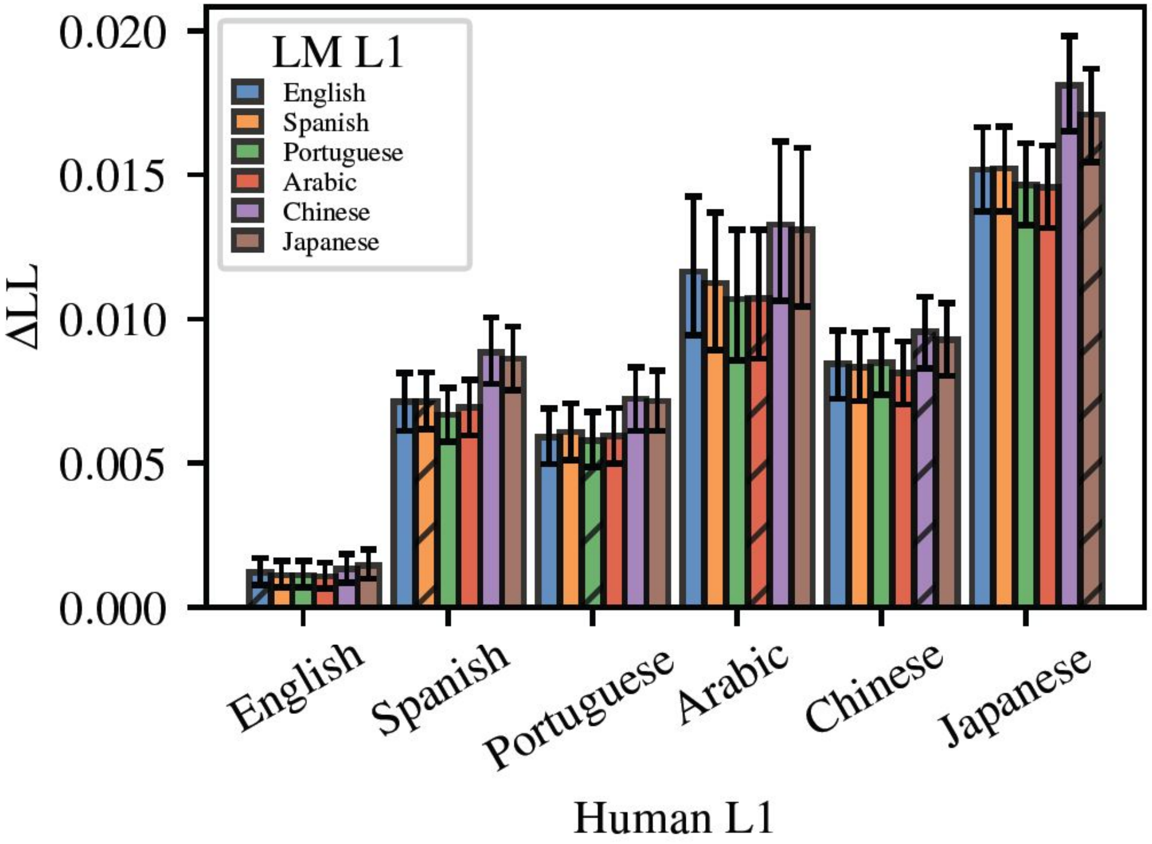
- This should apply to other languages as well

RQ2



- If our hypothesis is right, the result plot should look like this!

RQ2



Hypothesis disconfirmed!

- But ΔLL does vary by LM L1 (i.e. choice of pretraining L1 affects LM L2 sentence processing)

Human L1 was a stronger predictor of ΔLL ($F=628.64$, $df=5$, $p < .001$) than LM L1 was ($F=16.67$, $df=5$, $p<.001$)

check our paper for interesting examples!

RQ2

- This was from the final checkpoint…
- How L2LMs' PPP change during pretraining?

RQ2

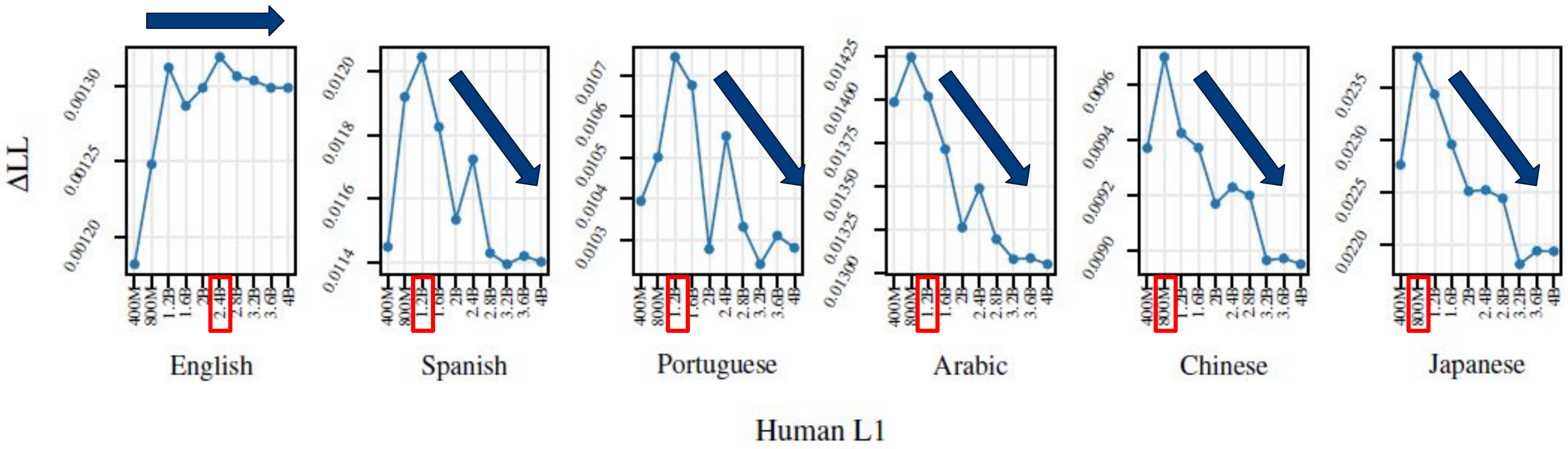


Figure 5: Reading time predictive power (ΔLL) of a monolingual English LM over the course of training. Plots reflect evaluation on data from different human L1 groups.

RQ2

x axis: L2 perplexity (lower the better)
y axis: L2 ΔLL (higher the better)

Hypothesized plot

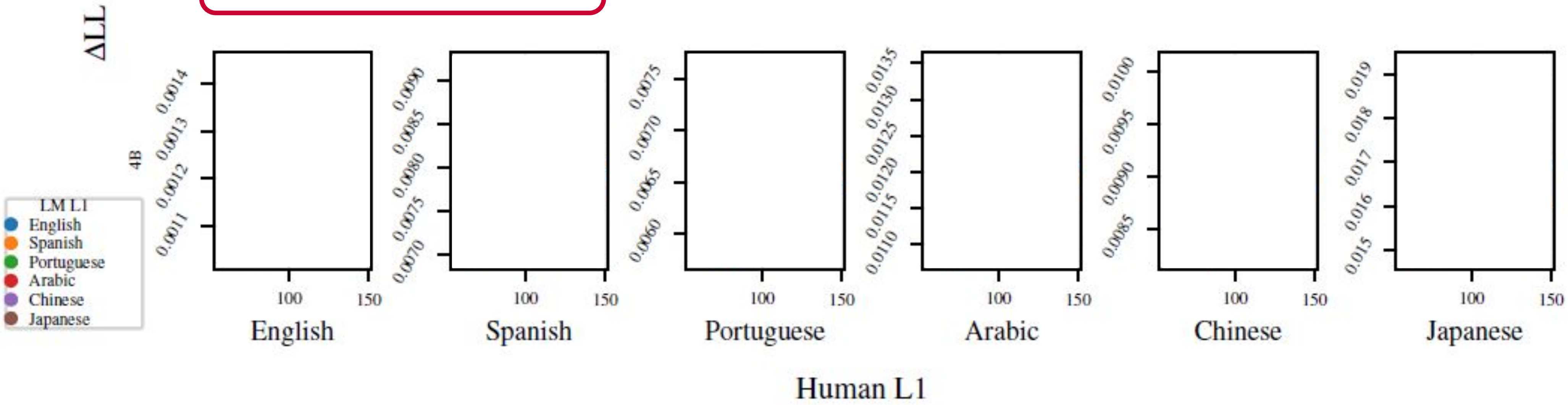


Figure 6: The relation between L2LMs' ΔLL (y-axis) and L2 perplexity (x-axis) at every 3M tokens during the L2 training phase. Each line represents an L2LM trained on the L1 of the corresponding color for 400M tokens (top) and 4B tokens (bottom), respectively. The shaded region around each line represents the 95% confidence interval. Human L1s are indicated on the x-axis.

RQ2

x axis: L2 perplexity (lower the better)
y axis: L2 ΔLL (higher the better)

Hypothesized plot

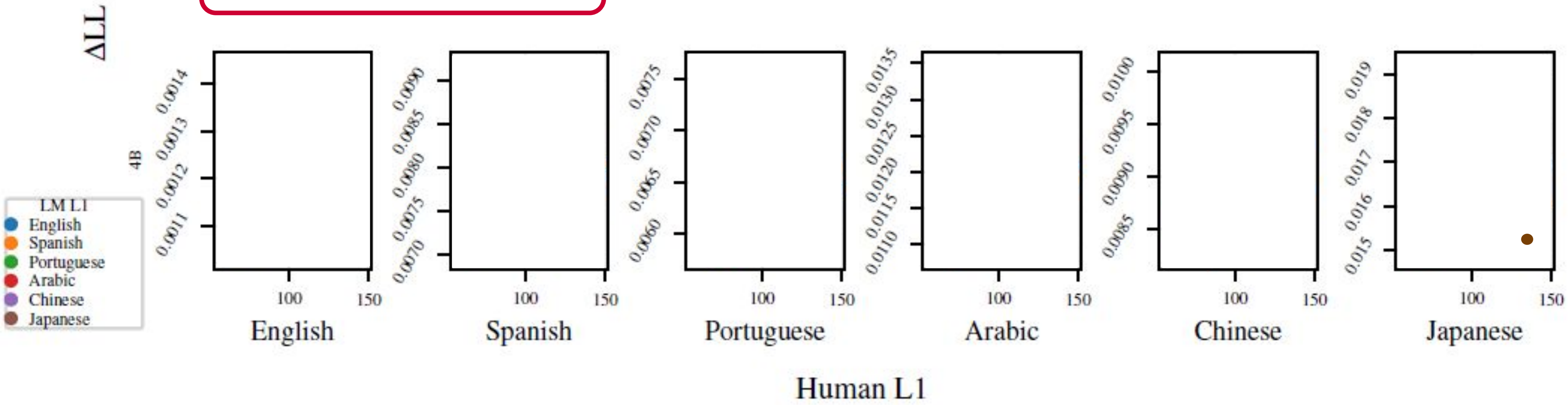


Figure 6: The relation between L2LMs' ΔLL (y-axis) and L2 perplexity (x-axis) at every 3M tokens during the L2 training phase. Each line represents an L2LM trained on the L1 of the corresponding color for 400M tokens (top) and 4B tokens (bottom), respectively. The shaded region around each line represents the 95% confidence interval. Human L1s are indicated on the x-axis.

RQ2

x axis: L2 perplexity (lower the better)
y axis: L2 ΔLL (higher the better)

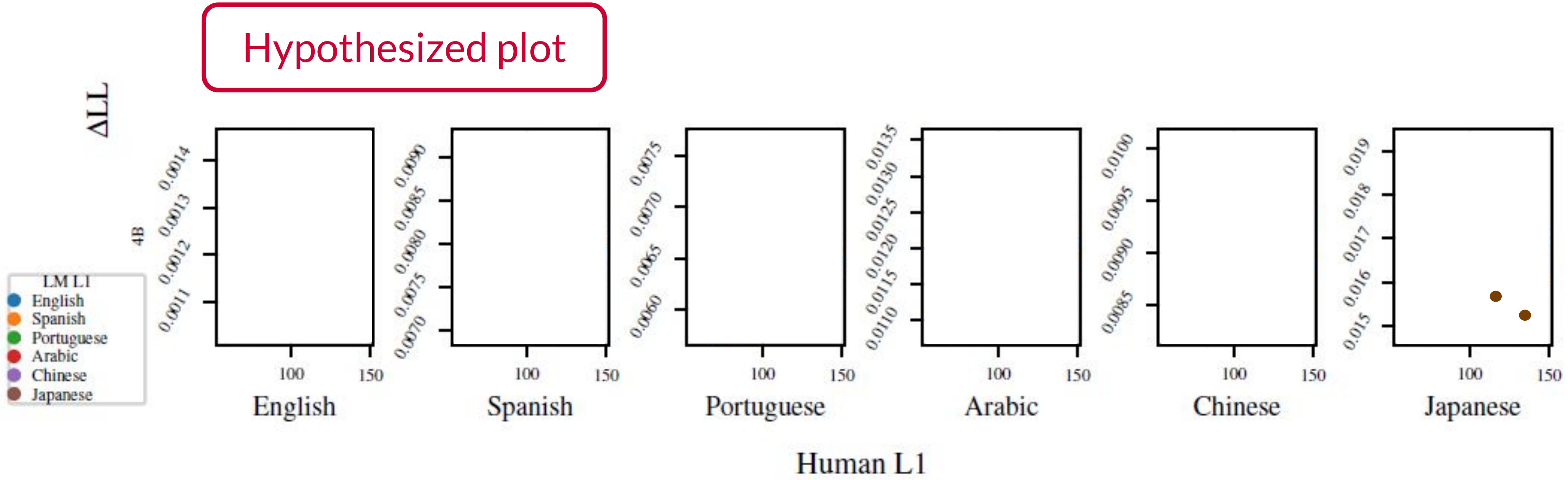


Figure 6: The relation between L2LMs' ΔLL (y-axis) and L2 perplexity (x-axis) at every 3M tokens during the L2 training phase. Each line represents an L2LM trained on the L1 of the corresponding color for 400M tokens (top) and 4B tokens (bottom), respectively. The shaded region around each line represents the 95% confidence interval. Human L1s are indicated on the x-axis.

RQ2

x axis: L2 perplexity (lower the better)
y axis: L2 ΔLL (higher the better)

Hypothesized plot

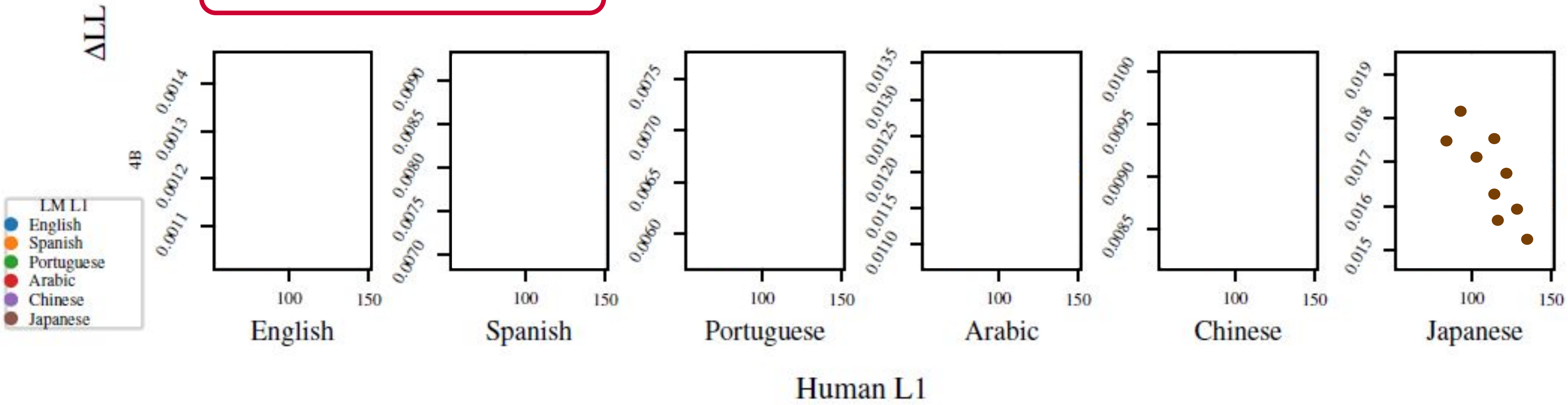


Figure 6: The relation between L2LMs' ΔLL (y-axis) and L2 perplexity (x-axis) at every 3M tokens during the L2 training phase. Each line represents an L2LM trained on the L1 of the corresponding color for 400M tokens (top) and 4B tokens (bottom), respectively. The shaded region around each line represents the 95% confidence interval. Human L1s are indicated on the x-axis.

RQ2

x axis: L2 perplexity (lower the better)
y axis: L2 ΔLL (higher the better)

Hypothesized plot

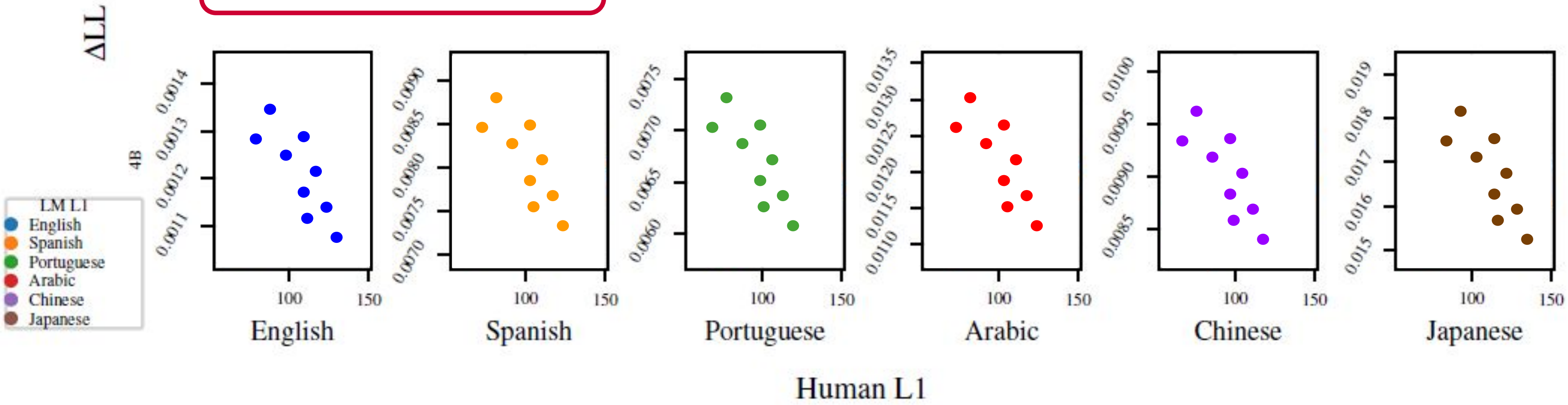


Figure 6: The relation between L2LMs' ΔLL (y-axis) and L2 perplexity (x-axis) at every 3M tokens during the L2 training phase. Each line represents an L2LM trained on the L1 of the corresponding color for 400M tokens (top) and 4B tokens (bottom), respectively. The shaded region around each line represents the 95% confidence interval. Human L1s are indicated on the x-axis.

RQ2

x axis: L2 perplexity (lower the better)
y axis: L2 ΔLL (higher the better)

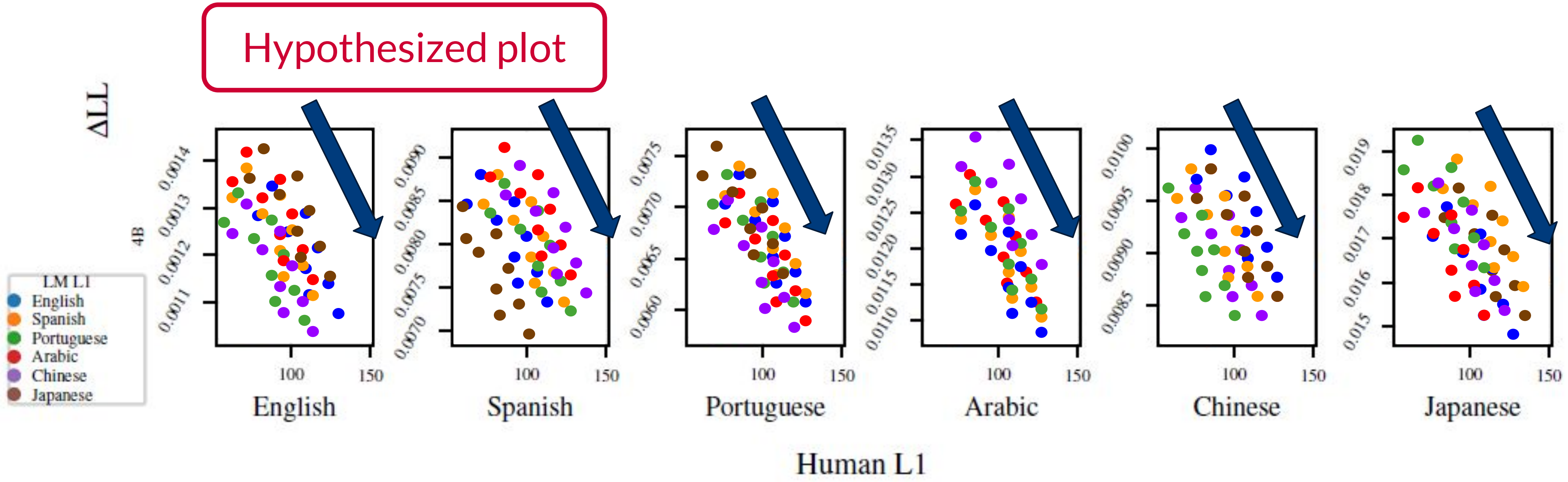


Figure 6: The relation between L2LMs' ΔLL (y-axis) and L2 perplexity (x-axis) at every 3M tokens during the L2 training phase. Each line represents an L2LM trained on the L1 of the corresponding color for 400M tokens (top) and 4B tokens (bottom), respectively. The shaded region around each line represents the 95% confidence interval. Human L1s are indicated on the x-axis.

RQ2

Throughout the L2 pretraining phase (3M-30M)
PPP and LM quality are in *negative* correlation!

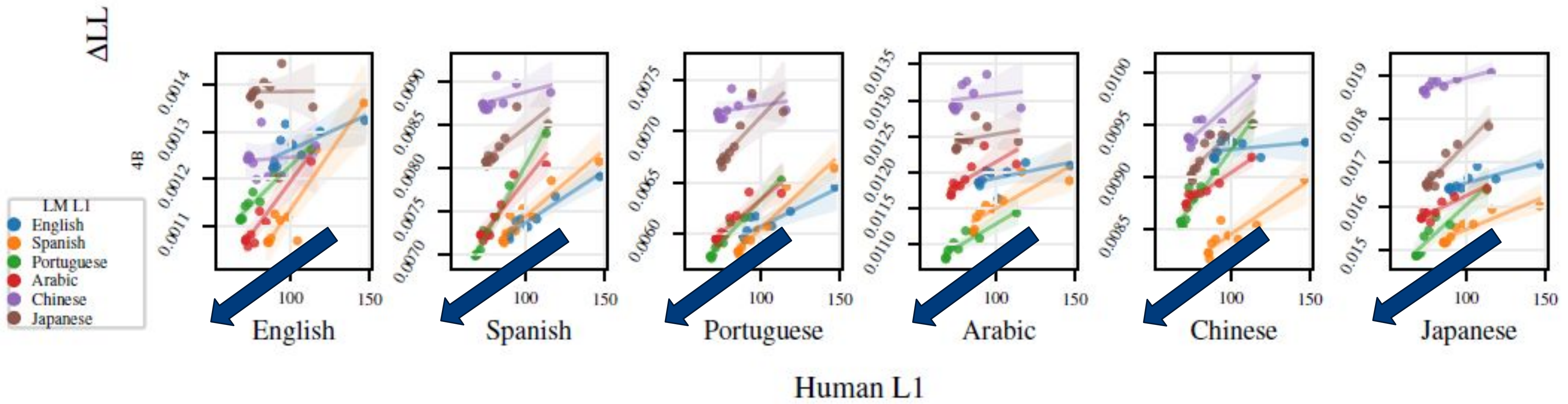
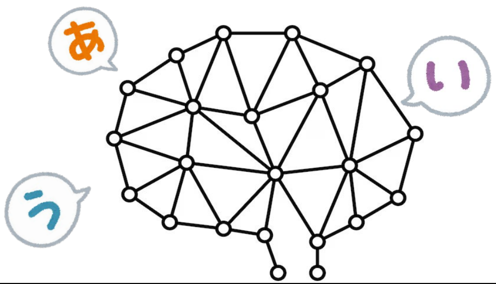


Figure 6: The relation between L2LMs' ΔLL (y-axis) and L2 perplexity (x-axis) at every 3M tokens during the L2 training phase. Each line represents an L2LM trained on the L1 of the corresponding color for 400M tokens (top) and 4B tokens (bottom), respectively. The shaded region around each line represents the 95% confidence interval. Human L1s are indicated on the x-axis.

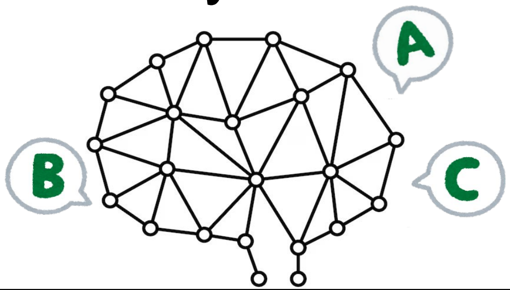
Takeaways

- LMs with different L1s read English differently



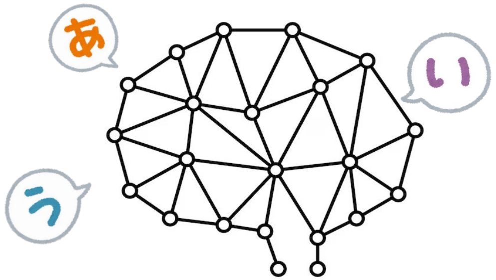
Pierre Vinken, 61 years old, will join…

≠



Pierre Vinken, 61 years old, will join…

- But not exactly like humans with the same L1!



Pierre Vinken, 61 years old, will join…

≠

Pierre Vinken, 61 years old, will join…

Takeaways

- L2 pretraining of up to 30M tokens lead to *lower* PPP
- Future direction
  - Why does L2 pretraining lead to *lower* PPP?
    (previously reported tipping point is 2B tokens; Oh & Schuler, 2023)
  - Establish a precise relationship between pretraining dynamics and PPP

Thank You!

References

- Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. CELER: A 365-Participant Corpus of Eye Movements in L1 and L2 English Reading. Open Mind, 6:41–50.
- Leshem Choshen, Ryan Cotterell, Michael Y. Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [Call for Papers] the 2nd BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. arXiv preprint 771 arXiv:2404.06214.
- Harald Clahsen and Claudia Felser. 2006. Continuity and shallow structures in language processing. Applied Psycholinguistics, 27(1):107–126.
- John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In Second Meeting of the North American Chapter of the Association for Computational Linguistics.
- Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In Proceedings of the 25th Conference on Computational Natural Language Learning, pages 624–646, Online. Association for Computational Linguistics.
- Roger Levy. 2008. Expectation-based syntactic comprehension. Cognition, 106(3):1126–1177.
- Akira Murakami and Theodora Alexopoulou. 2016. L1 influence on the acquisition order of english grammatical morphemes: A learner corpus study. Studies in Second Language Acquisition, 38(3):365–401.
- Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.
- Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1915–1921, Singapore. Association for Computational Linguistics.
- Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6829–6839, Online. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. Transactions of the Association for Computational Linguistics, 8:377–392.
- Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023. Language model quality correlates with psychometric predictive power in multiple languages. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7503–7511, Singapore. Association for Computational Linguistics.
- Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real time comprehension behavior. In Proceedings of the 42nd Annual Meeting of the Cognitive Science Society, page 1707–1713.
- Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.