Can large language models predict English phrasal stress?

Jinyoung Jo[1], Sean Choi[2], and Arto Anttila[1]

[1]Stanford University and [2]Santa Clara University

AMP 2025, Special Session on Deep Phonology

UC Berkeley
September 27, 2025

Sentential prominence

- Words in English sentences show degrees of prominence:

And this is my solemn pledge

George W. Bush's first inaugural, January 20, 2001, transcribed by a native speaker, degrees of prominence (0, 1, 0, 2, 2, 3) visualized by font size

- What explains these degrees of prominence?

Two key predictors of sentential prominence

- **1. Mechanical stress:** Syntax and lexical stress predict phrasal and sentential prominence to a good approximation.
  - Nuclear/Compound Stress Rules (NSR/CSR, Chomsky et al. 1956; Chomsky and Halle 1968; Liberman and Prince 1977; Cinque 1993)
  - Alternative: The pitch accent view [not addressed in this talk] (Gussenhoven 2011)

Two key predictors of sentential prominence

- **2. Meaningful stress:** Informative (i.e., new, highlighted, focused) words are prominent (Bolinger 1972). Informativity can be approximated in various ways:
  - Average predictability (Cohen Priva 2015):
    Informativity in cold storage (≈ the lexicon) that does not vary across contexts
  - Contextual probability from LLMs:
    Informativity computed on the fly using the context
    - Logits (unnormalized, raw scores)
    - Log probabilities (log of normalized probabilities)

The Present Study

LLMs are good at predicting the next word, which should help quantify the word's informativity, and by hypothesis, predict its prominence.

- Do LLM probabilities help predict a word's prominence?
- Do LLMs improve on NSR/CSR and average predictability?

Methods: Data

- **The Presidents Project**: An ongoing project on the prosody of presidential speeches, annotated by humans and machines
  (Shapiro 2019; Anttila et al. 2020; Anttila and Shapiro 2020; Clapp and Anttila 2021; Shapiro and Anttila 2021)
  - This talk: The first inaugurals of Bush (2001) and Obama (2009)
    - 21,686 data points
  - Annotated for syntax, phonology, NSR/CSR, informativity, etc.
  - Annotated for perceived prominence by 7 native speakers
    - 8-point scale (1 least prominent, 8 most prominent)

Methods: Predictors of Prominence

1. The Nuclear/Compound Stress Rule (NSR/CSR)
   - A version of NSR/CSR is implemented in MetricalTree `https://metricaltree.stanford.edu/` (Anttila et al. 2020)
   - The algorithm builds on syntax from the Stanford Parser (Klein and Manning 2003; Chen and Manning 2014; Manning et al. 2014)

[[[John's]	[[[black]	[board]]	[eraser]]]	[was stolen]]	
1	1	1	1	1	Lexical stresses
	[1	2]			Cycle 1 (CSR)
	[1	3	2]		Cycle 2 (CSR)
[2	1	4	3]		Cycle 3 (NSR)
[3	2	5	4	1]	Cycle 4 (NSR)

   - 1 = primary, 2 = secondary, etc. Bigger number, less stress.

Methods: Predictors of Prominence

2. Informativity (Cohen Priva 2015)
   - Weighted average of the negative log predictability of seeing word w given each context c that w follows in the corpus

$$\text{inform}(w) = -\sum_{c \in C} P(c \mid w) \log_2 P(w \mid c)$$

   - Informativity was added to the corpus by Naomi Shapiro. We used bigram informativity.
   - Bigger number, more information, more stress.

Methods: Predictors of Prominence

3. Contextual probability
   - Llama 3.2 3B (Touvron et al. 2023)
     - We asked the model to calculate log probabilities of candidate next words given context, then retrieved the target word's log probability.
     - e.g. Obama (2009) begins with *my fellow citizens*...

Prompt (context)	Candidate next word	Log probability
my	name	-2.44
	friend	-3.36
	...	...
	fellow	-7.33
	...	...
my fellow	Americans	-1.60
	citizens	-1.68
	students	-3.90
	...	...
...	...	...

Methods: Predictors of Prominence

3. Contextual probability
   - Bigger number, higher probability, less stress.
   - Rare words not returned by the model were excluded (2.7% of data).

Methods: Regression Models

- Mixed effects linear regression using *lmer()* in *lme4* (Bates et al. 2015):

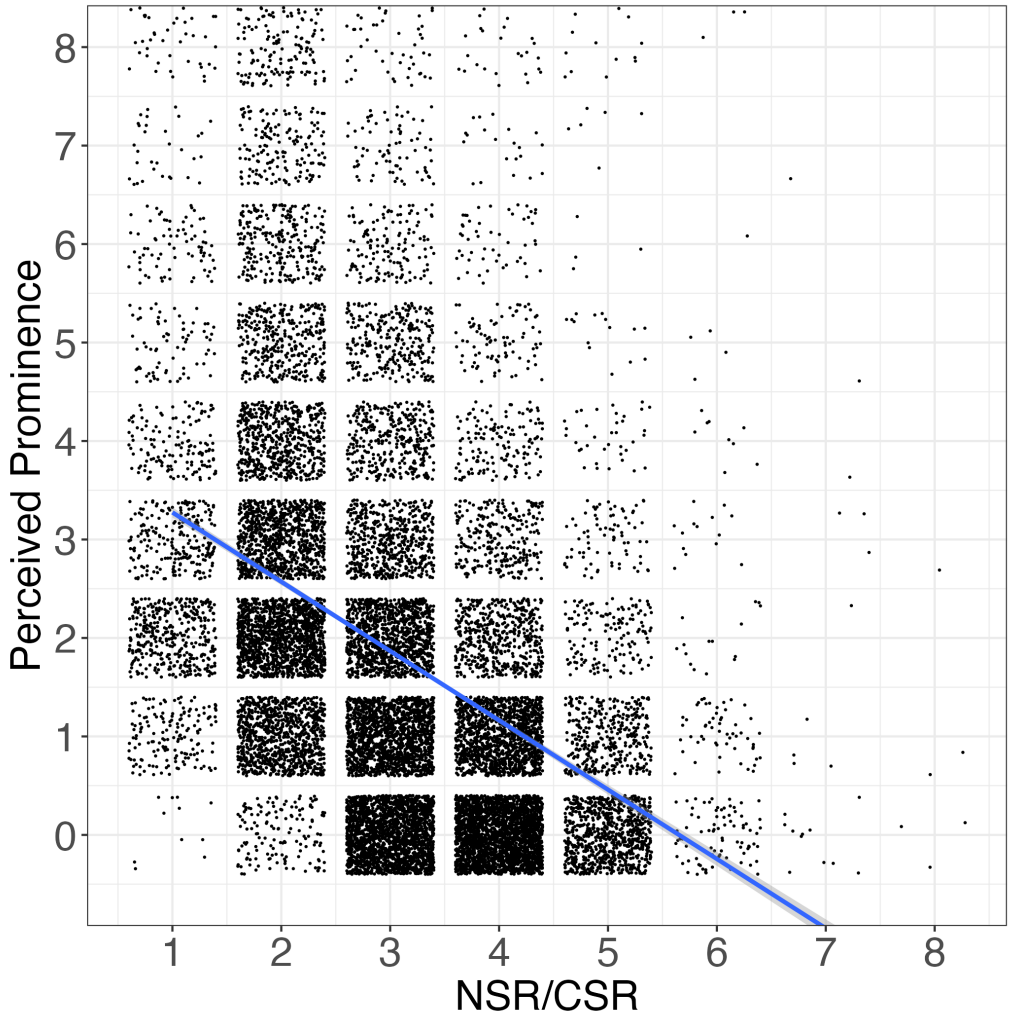
  Perceived Prominence ~ NSR/CSR + Bigram Informativity + Log Probability + (1|Annotator)

- All predictors were scaled for comparability.

- Consistent results were obtained in an ordinal logistic regression model.

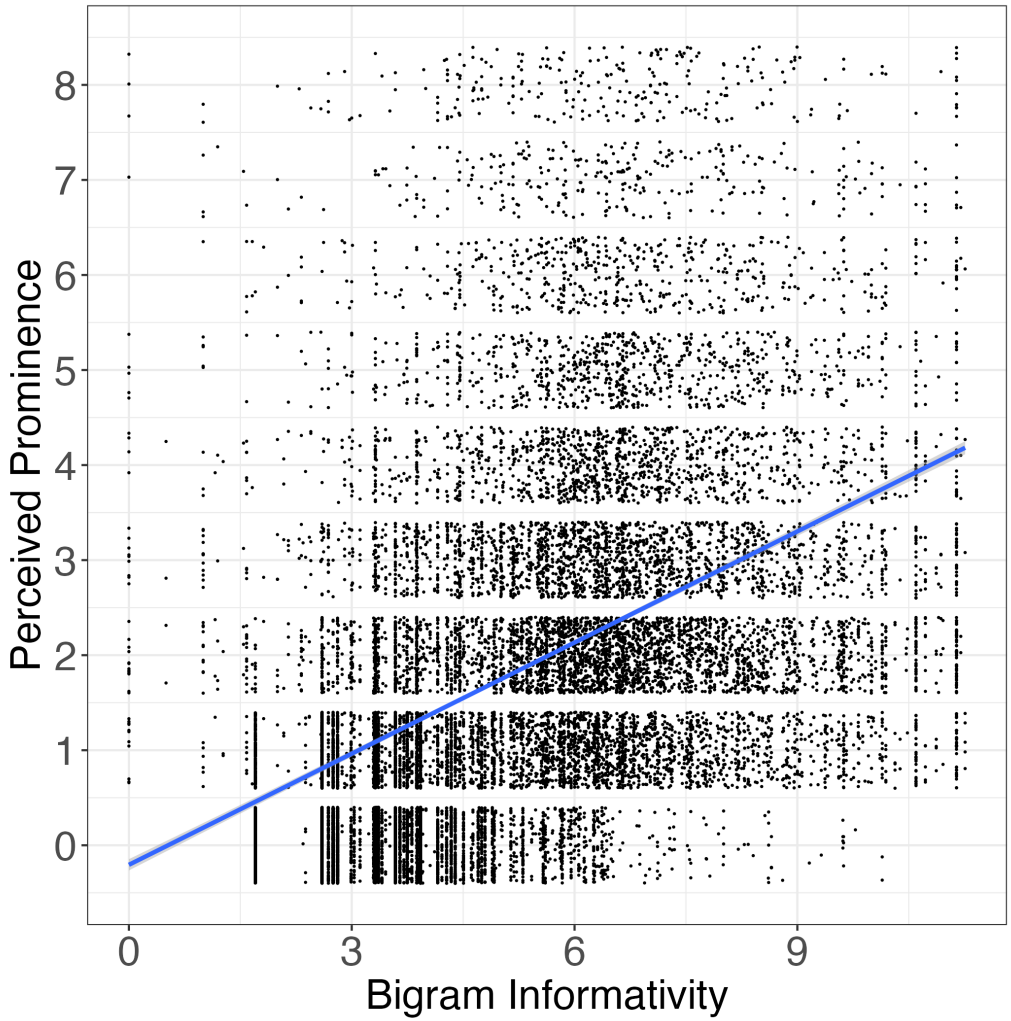
Results for NSR/CSR (mechanical stress)

- NSR/CSR is negatively correlated with perceived prominence (r=-0.41, p<0.001) as expected.



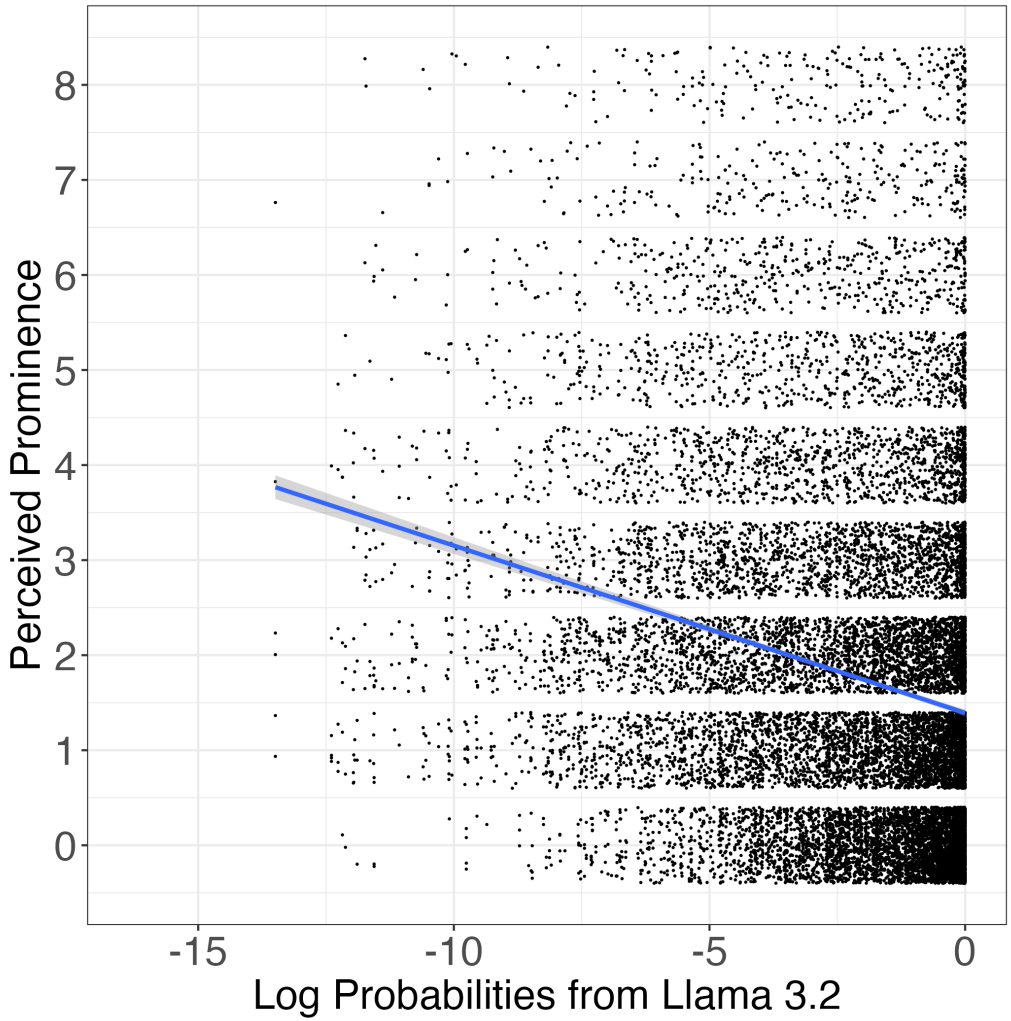
Results for informativity (meaningful stress)

- Bigram informativity is positively correlated with perceived prominence (r=0.47, p<0.001) as expected.

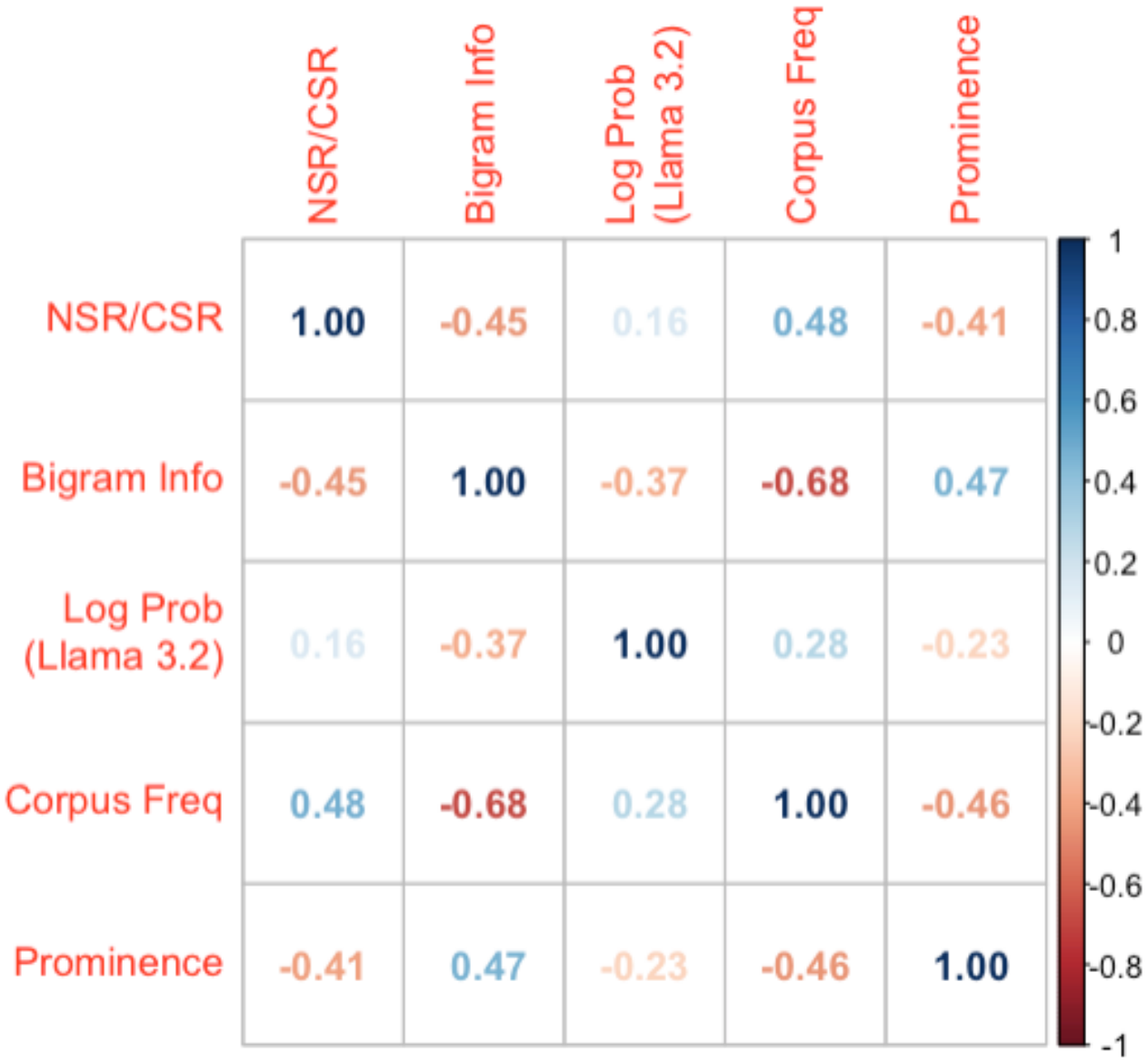


Results for LLM (meaningful stress)

- Log probability is negatively correlated with perceived prominence (r=-0.23, p<0.001) as expected, but more weakly.



Correlation matrix



Mixed effects linear regression model

- Controlling for other predictors, LLMs help predict perceived prominence, but the effect is smallish.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.11	0.43	4.95	< 0.001
NSR/CSR	-0.47	0.01	-45.60	< 0.001
Bigram Informativity	0.65	0.01	57.51	< 0.001
Log Probability	-0.11	0.01	-10.24	< 0.001

Ongoing Work

- We are expanding our data set ($\sim$ 80k data points).
- Consistent results were obtained with two other models:
  - Llama 2 13B
  - Mistral 7B

Conclusions

- The effect of contextual predictability operationalized through LLMs is real but remains relatively small.
- The production and perception of phrasal prominence has
  - a syntactic basis (NSR/CSR)
  - a lexical basis (word stress, average predictability)

  that is context-independent.

Acknowledgements

Funding: This research builds on work funded by the Roberta Bowman Denning Initiative in the Digital Humanities as part of the project *Prose Rhythm and Linguistic Theory*, the Vice-Provost for Undergraduate Education at Stanford University, and Stanford Introductory Seminars Plus.

People: We thank Uriel Cohen Priva and Naomi Shapiro for suggesting the idea that resulted in this study. Other individuals who contributed include Frankie Conover, Timothy Dozat, Elena Felix, Vivienne Fong, Daniel Galbraith, Julia Mendelsohn, Shina Penaranda, Liam Smith, Madeline Snigaroff, Saahil Sundaresan, Connor Toups, Alexander Wade, Amy Wang, Annalisa Welinder, and Amir Zur.

Data: The inaugurals data were obtained from Peters and Woolley (1999–) *The American Presidency Project*, `http://www.presidency.ucsb.edu/index.php`.

Thank you!

References I

Anttila, Arto, Timothy Dozat, Daniel Galbraith, and Naomi Shapiro. 2020. Sentence stress in presidential speeches. *Prosody in Syntactic Encoding, Berlin/Boston: Walter De Gruyter* 17–50.

Anttila, Arto, and Naomi Shapiro. 2020. Studying sentence stress using corpus data. A handout from a workshop on intonation, *Annual Meeting on Phonology 2020*, UC Santa Cruz, `https://web.stanford.edu/~anttila/research/AMP-2020-Tutorial-Handout.pdf`.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1–48.

Bolinger, Dwight L. 1972. Accent is predictable (if you are a mind reader). *Language* 48:633–644.

Chen, Danqi, and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 740–750.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. Cambridge, Mass.: MIT Press.

Chomsky, Noam, Morris Halle, and Fred Lukoff. 1956. On accent and juncture in English. In *For Roman Jakobson: Essays on the occasion of his sixtieth birthday*, 65–80. The Hague: Mouton Co.

Cinque, Guglielmo. 1993. A null theory of phrase and compound stress. *Linguistic Inquiry* 24:239–298.

References II

Clapp, William, and Arto Anttila. 2021. To predict or to memorize: Prominence in inaugural addresses. In *Supplemental Proceedings of the 2020 Annual Meeting on Phonology*. Washington, DC: Linguistic Society of America.

Cohen Priva, Uriel. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6:243Äì278.

Gussenhoven, Carlos. 2011. Sentential prominence in English. In *The Blackwell Companion to Phonology*, ed. Marc Oostendorp, Colin J. Ewen, Elizabeth Hume, and Keren Rice, 2778–2806. Malden, MA: Wiley-Blackwell.

Klein, Dan, and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, 423–430.

Liberman, Mark, and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8:249–336.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford coreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.

Shapiro, Naomi Tachikawa. 2019. MetricGold. A software package for the metrical annotation of English prosody, `https://github.com/tsnaomi/metric-gold`.

References III

Shapiro, Naomi Tachikawa, and Arto Anttila. 2021. On the phonology and semantics of deaccentuation. In *Proceedings of the 2020 Annual Meeting on Phonology*, ed. Ryan Bennett, Richard Bibbs, Mykel L. Brinkerhoff, Max J. Kaplan, Stephanie Rich, Amanda Rysling, Nicholas Van Handel, and Maya Wax Cavallaro. Washington, DC: Linguistic Society of America.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* .