

CHISEL: Chunk Holdout Influence Scoring by Elimination Loss

Removing pieces to reveal what matters

Dan Petrovic
DEJAN, dejan.ai
October 2026

ABSTRACT

AI search systems answer a prompt by citing a few web sources, each represented to the model by a short grounding snippet. We ask which parts of a cited source's text earn its rank among the other cited sources. Removing one part at a time does not answer this: across 1,065 single-line removals recorded in our runs, 60% left the source's rank unchanged, including 103 lines whose measured contribution is clearly non-zero. CHISEL treats the spans of a source's text, plus its brand, as players in a cooperative game whose value is the negative of the source's rank under a large language model (LLM) ranker, and estimates each player's Shapley value by antithetic permutation sampling with coalition caching, a precision stop, a budget cap and a 30-player limit that groups long texts into topic blocks. Over 111 runs on 47 prompts across Google, ChatGPT and Claude, each run's values add up to its rank gain to within 0.002 places on average, and 278 of 288 clearly non-zero values keep their sign when a run's rankings are split in half. Lines that name the product or topic of the prompt are worth 1.7 to 2.5 times an average line, the opening line 1.4 to 3.8 times, and page furniture such as navigation and captions close to nothing; the source's brand name moves its rank by less than an eighth of a place on average.

Keywords: AI search, grounding snippets, Shapley values, feature attribution, LLM ranking, occlusion

1. Introduction

When an AI search system grounds an answer in web results, it reads a short selection of text from each retrieved page, a grounding snippet, and cites a subset of those pages [1]. For a publisher, the snippet is the unit that competes. Knowing which sentences of a cited snippet earn its position tells a writer what their own page needs to say.

The obvious experiment is to delete one sentence, signal or brand mention, rank the sources again and record the change. We ran this many times and it returned no effect in most cases (Section 3). The cause is known in the attribution literature: when several parts of an input carry the same information, deleting any one leaves the rest to carry it, so each deletion shows zero influence even when the parts together decide the outcome [2, 3]. A rank is also an integer with a floor, and a source that ranks first has nowhere to go until enough of its support is gone.

CHISEL replaces single deletions with Shapley attribution over fixed chunks of text. Each chunk is held out across many combinations of the others, and its score is the average rank lost when it is eliminated. The contributions are:

- evidence from 1,065 single-line removals that deletions one at a time miss most effects;
- the CHISEL procedure, including a 30-player limit with model-drawn topic blocks for long texts;
- validity checks on 111 runs: additivity, split-half agreement and an edit test;
- what CHISEL finds across three platforms, with worked examples.

2. Problem setting

A prompt q is run on an AI platform from a chosen country with a short system instruction to answer as a recommendation engine in a single turn, which stops the platform from replying with a clarifying question. The platform returns an answer and the sources it cited, $C = \{c_1, \dots, c_m\}$. Each source has a

text x_c : for Google, the snippet the model reproduces through a dump prompt; for OpenAI, the snippet text its API returns; for Anthropic, whose API returns the text in encrypted form, an extractive snippet written from the fetched page. When a platform returns no text for a cited source, an extractive snippet is written for that source. The target t is by default the source the platform cited first.

The platform's ranking function is not observable, so a ranker R stands in for it: an LLM given the prompt and the labelled texts of all cited sources, which returns a full ordering. Each platform's runs used that platform's model as R (Gemini 3.8 Flash, GPT-5.6 and Claude Sonnet 5.5). On 54 saved Google grounding runs, the median of three rankings by R matched Google's citation order with mean Spearman correlation 0.40 on 30 development prompts and 0.34 on 24 held-out prompts, and matched Google's first-cited source in 22 of 54. CHISEL therefore measures what moves the ranker; Section 7 returns to this.

3. Single deletions find little

Early experiments. Before CHISEL we tested deletion directly on one agency-services prompt. In five runs, a model listed the target's signals as exact spans and each was deleted in turn, with three rankings per deletion: 87 of 88 deletions left the target's median rank unchanged. Deleting every mention of the target's brand, with the sources rotated through every input position, left it first in 68 of 72 rankings. Deleting each of its nine sentences in turn left it first in all 30 rankings, with or without the brand.

At scale. CHISEL's own rankings include every coalition that is missing exactly one player whenever an order ends with that player, which gives single-removal measurements for free. Across the 111 runs, 1,065 such removals were recorded. 635 (60%) left the target's rank unchanged. Of those 635 lines, 103 have a Shapley value more than twice its standard error away from zero (Figure 1). The share of unchanged removals was 69% on Google, 71% on ChatGPT and 30% on Claude, whose answers cite ten sources and whose ranks move more easily.

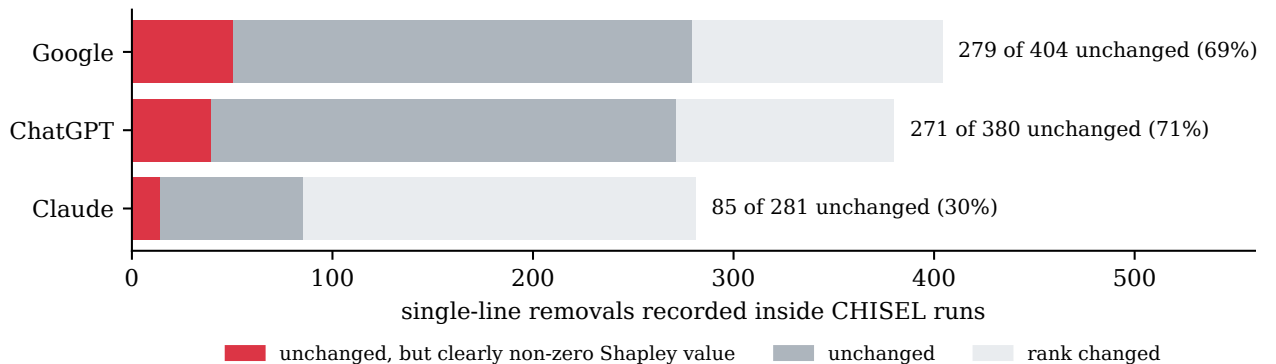


Figure 1. Single-line removals recorded inside CHISEL runs, by platform.

Worked example. owayo.com ranks first among five sources for "custom running shirts" on Google. Removing any one of its five sentences, or its brand name, leaves it first. CHISEL, with 62 rankings, divides the four places between its full and empty text among the same lines (Figure 2): the sentence that lists the shirt types ($+1.55 \pm 0.25$), "owayo manufactures custom running shirts of professional quality." ($+1.27 \pm 0.22$) and the title "Custom Running Shirts." ($+0.66 \pm 0.18$). The brand name holds $+0.04 \pm 0.05$.

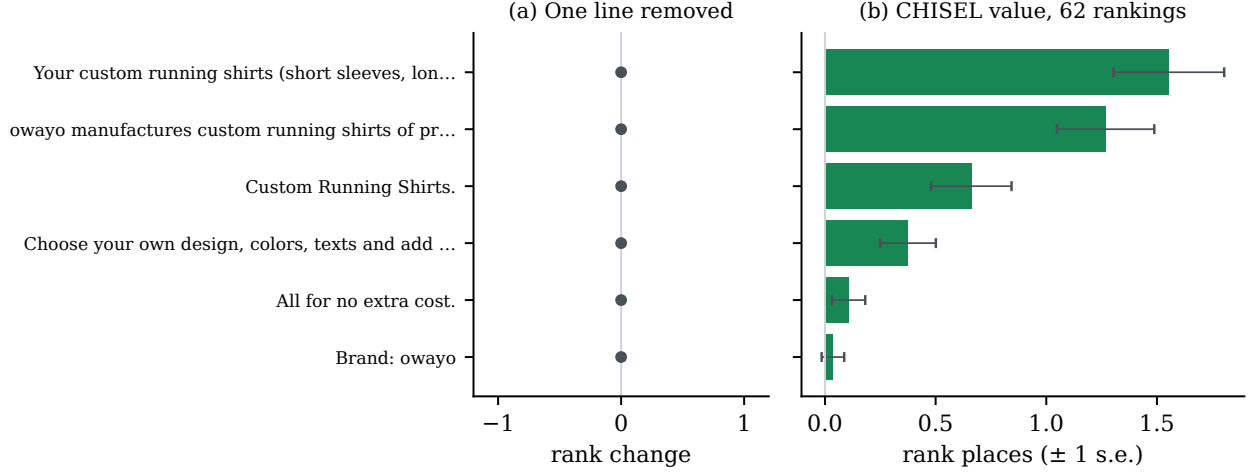


Figure 2. owayo.com for "custom running shirts" on Google (owayo.com/custom-running-shirts.htm). (a) Removing any single line leaves the rank unchanged. (b) CHISEL's Shapley values for the same lines. Error bars are one standard error.

4. Method

4.1 Players

The target's text is cut into spans that join back into the original character for character: sentences, or the parts between the "..." separators Google places between extracts, in Snippet mode; paragraphs (with list items, table rows and headings as their own spans) or sentences in Page mode. A sentence that occurs more than once is one player covering every copy. The brand is one further player: its pattern matches the site's domain and the domain's first label with optional spaces or hyphens between its letters, and when it is absent from a coalition every match is deleted wherever its sentence is still present.

A run has at most 30 players. When the spans and the brand would exceed 30, a model receives the spans numbered and groups consecutive spans into exactly as many blocks as fit, of similar word count and with boundaries at topic changes; it returns only the start of each block. The code cuts the text at those points, so the text is never rewritten, and falls back to blocks of equal word count if the reply fails its checks. 25 of the 111 runs were grouped this way.

4.2 Coalition texts and the value function

For a coalition $S \subseteq N$, the text x_S keeps the spans in S in order. Only the seam a removal leaves is tidied, and a "..." separator is kept only with text on both sides. The value of a coalition is

$$v(S) = -\text{rank}_R(t \mid x_S),$$

measured with one ranking. In Snippet mode each ranking draws a fresh random assignment of labels and a fresh input order; in Page mode the target is listed last and labels are fixed, which keeps the long shared prefix cacheable.

4.3 Shapley values

The Shapley value [4] of player i is its average marginal contribution over all orders of adding the players:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} |S|! (n - |S| - 1)! / n! \cdot [v(S \cup \{i\}) - v(S)].$$

With v defined by rank, φ_i is the average number of rank places the target gains when span i is present: the elimination loss of the span. The values add up to $\text{rank}_R(t \mid \emptyset) - \text{rank}_R(t \mid x_t)$. CHISEL estimates φ by permutation sampling [5, 6]: each round draws random orders and also walks each in reverse, an

antithetic pair [7]; every coalition is ranked at most once and cached; and when all 2^n coalitions have been ranked the values are computed exactly.

4.4 Stopping and budget

A run stops when at least eight orders have been walked and every standard error is at most 0.25 places, when the next round would pass its budget, or when every coalition has been ranked. The default budget is \$20 at retail prices. Usage is billed after every round, and an interrupted run can resume from its saved rankings.

Algorithm 1 CHISEL

```

input: prompt q, cited texts {x_c}, target t, ranker R, players N (at most 30), tau, budget B
cache <- {}; orders <- []
repeat
  if |cache| = 2^n: return exact Shapley values from cache
  draw k random orders; add each order and its reverse to the round
  need <- coalitions on the round's orders not in cache
  if spent + mean_cost * |need| > B: stop (budget)
  for S in need, in parallel:
    cache[S] <- -rank_R(t | x_S)      # one ranking, labels and order shuffled
  orders <- orders + round
  for each player i: phi_i <- mean of cache[S+i] - cache[S] over walked orders
                    se_i <- sd / sqrt(count)
until |orders| >= 8 and max_i se_i <= tau
return phi, se

```

5. Experiments

5.1 Data

The analysis uses every completed run: 111 runs on 47 prompts in 11 projects, with 2,178 players. 82 runs are product prompts from one sportswear manufacturer's catalogue run on all three platforms; the rest cover software, travel, finance, real-estate, retail and home-improvement prompts. Platforms differ in how much text they hand to the model and how many sources they cite (Table 1).

Platform	Runs	Settled	Stopped at budget	Sources cited	Snippet words	Players	Rankings	Retail price
Google	53	47	6	5	53	8	88	\$0.84
ChatGPT	33	17	16	6	227	29	458	\$17.00
Claude	25	1	24	10	340	26	364	\$19.46

Table 1. Runs by platform. "Settled" counts runs that met the precision rule or ranked every combination. Sources, snippet words (per cited source; Claude's are extractive snippets written from the page), players, rankings and price are medians per run.

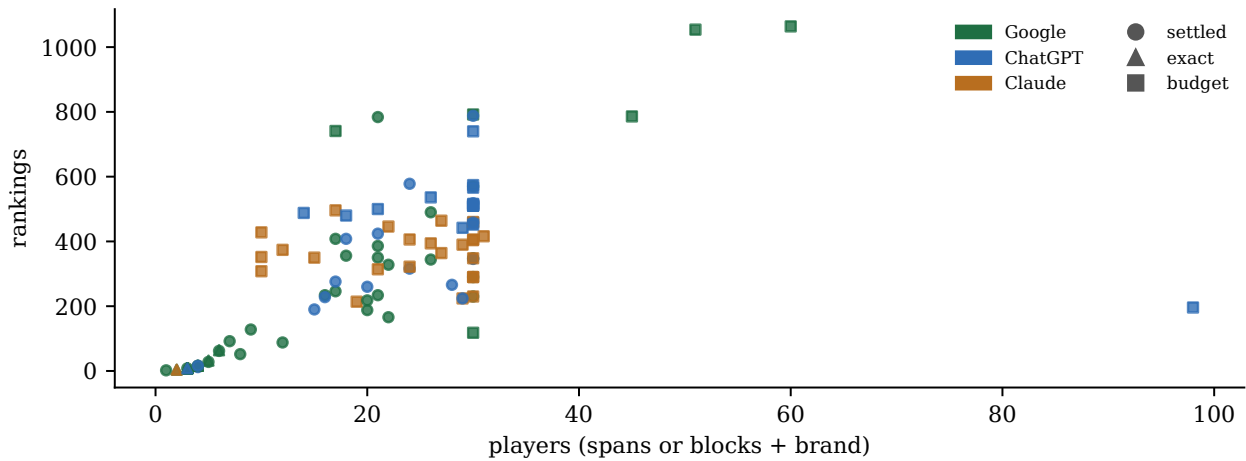


Figure 3. Rankings used against the number of players, by platform and stop reason. The five runs above 30 players predate the 30-player limit.

Google snippets are short, so most Google runs settle within a few dollars and 19 ranked every combination. ChatGPT and Claude hand the model longer texts from more sources; most of their runs reached the budget first. Their values are kept and reported as rough: a fixed threshold of 0.25 places ignores that Claude ranks ten sources where Google ranks five.

5.2 Validity checks

Additivity. In 110 runs with a ranking for both the empty and the full text, the values sum to the rank gain to within 0.0015 places on average and 0.053 at most.

Split-half agreement. Each run's walked orders were split at random into two halves, keeping antithetic pairs together, and the values recomputed from each half (Table 2). Values clearly different from zero keep their sign in 97% of cases on every platform. The ordering of the top spans is stable on Google and ChatGPT and unstable on Claude.

Platform	Runs	Same sign	Top span in other half's top 3	Rank correlation
Google	36	103 of 106	24 of 36	0.41
ChatGPT	30	107 of 110	24 of 30	0.37
Claude	24	68 of 72	7 of 24	0.19

Table 2. Split-half agreement, for runs with at least four players and six orders. "Same sign": values more than two standard errors from zero in the full run that keep their sign in both halves. "Rank correlation": median Spearman correlation of all values between halves. Rank correlations are low because most spans have values near zero, where ordering is noise.

Edit test. In the early agency-services experiment, the two top-valued lines of the target, with each competitor's own name substituted, were added to the start of one competitor's text at a time, and each set was ranked five times. No competitor ranked first in any ranking before the edit. After it, four of the five ranked first at least once and two improved their median rank; the largest move, from median 5 to 1, came from a competitor whose name is itself the prompt.

5.3 What moves the rank

Spans differ in length and runs in how many places the target gains, so effects are reported relative to an even share of each run's gain: 1.0 is an average line of its page, 2.0 twice that and 0 nothing. Every span was labelled by an LLM with one of ten content types and with whether it names the product or topic of the prompt.

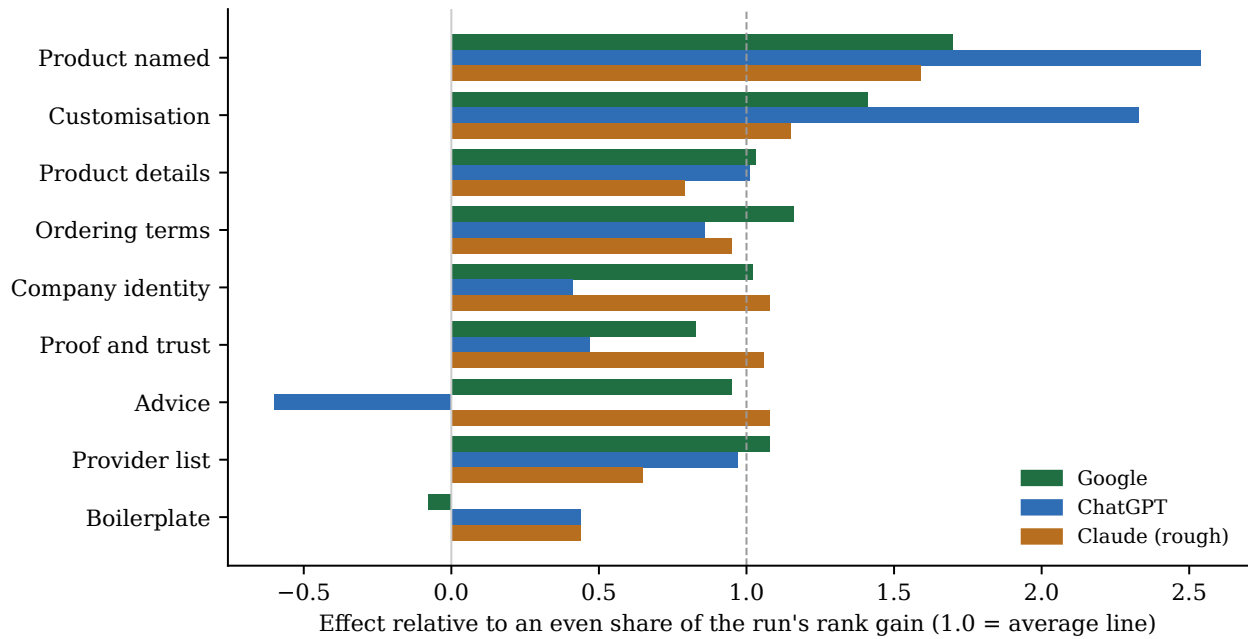


Figure 4. Effect of each content type relative to an average line of the same page, for types with at least five lines on a platform.

- **Naming the prompt's product or topic.** Lines that name it are worth 2.4 times an average line on Google, 2.5 on ChatGPT and 1.7 on Claude; lines that do not, 0.6 to 0.7. It is the highest-valued content type on all three platforms.
- **The opening line** is worth 1.4 times an average line on Google, 3.3 on ChatGPT and 3.8 on Claude.
- **Customisation and product capability** lines rank second on all three platforms (1.4, 2.3 and 1.2) and were the most valuable line of the page in 31 of 111 runs, more than any other type.
- **Page furniture** (navigation, buttons, captions) is worth -0.08 on Google and 0.44 on ChatGPT and Claude.
- **The brand name**, every mention tested together, averages $+0.12$ places on Google (30 runs), $+0.06$ on ChatGPT (23) and -0.05 on Claude (18).
- **Concentration.** The most valuable lines making up a median 47% of a page's words hold 80% of its gain on Google, 42% on ChatGPT and 39% on Claude.

The content types reflect a dataset in which three quarters of the runs are product prompts; other kinds of prompt may reward different content.

5.4 A concentrated case

For "Are there trackers that compare desktop versus mobile Google search volatility?" on Google from Australia (run 33), the source cited first was a semrush.com article among eight sources. It moved from rank 8 with no text to rank 3 with its full text, and one sentence, "Semrush Sensor measures volatility in search results, tracking down 20+ categories on mobile...", held $+4.10 \pm 0.10$ of the five places. The other 19 players held between -0.20 and $+0.20$ each. A second run in the same project (run 18, "google algorithm volatility tracking tool") gave its whole one-place gain to one sentence that names the topic: "Below are 5 algorithm monitoring tools that measure search engine volatility..." ($+1.04 \pm 0.21$).

6. Related work

Deleting parts of an input to measure their effect is occlusion in vision [8] and representation erasure in text [2]. Removal-based explanations, including leave-one-out, LIME [9] and SHAP [3], fit one framework that differs in how features are removed and how effects are summarised [10]; CHISEL removes spans by deletion and summarises with Shapley values. Sampling estimators for Shapley values go back to Castro et al. [5] and Štrumbelj and Kononenko [6], and antithetic permutations are

among the variance reduction methods compared by Mitchell et al. [7]. Rationale faithfulness in NLP is often tested by deleting the rationale and measuring the output change [11]; the edit test inverts this by adding top-valued spans elsewhere. Listwise reranking with LLMs [12] supplies the ranker, and LLMs weigh input positions unevenly [13], which is why CHISEL shuffles labels and order for every ranking.

7. Limitations

- **The ranker is a proxy.** CHISEL measures what moves R , which agreed with Google's citation order at a Spearman correlation of 0.34 on held-out prompts; Google's own order also varies between runs of the same prompt.
- **One ranking per coalition.** Each coalition's value is a single ranking, cached and reused across orders. Standard errors treat marginal contributions as independent although orders share coalitions, so they are optimistic.
- **Fixed precision threshold.** The 0.25-place stop ignores the number of sources ranked; 40 of 58 ChatGPT and Claude runs reached the budget first, and their values are reported as rough.
- **Integer outcome.** Rank moves in whole places; a continuous score such as the probability of being ranked first would separate small effects with fewer rankings.
- **Content labels** come from an LLM and the dataset is dominated by product prompts from one catalogue.
- **Claude snippets** are written by our extractor, since Claude's API does not expose the text it read.

8. Availability

CHISEL runs in the [DEJAN app](https://dejan.ai) under Content Optimizer → Ablation, open to all users. Runs 18 and 33 are in the public demo project. The feature is described at dejan.ai/blog/ablation and the method at dejan.ai/concepts/grounding-snippet-ablation.

References

1. D. Petrovic. Grounding snippets. dejan.ai, 2026. dejan.ai/concepts/grounding-snippet/
2. J. Li, W. Monroe, D. Jurafsky. Understanding neural networks through representation erasure. arXiv:1612.08220, 2016. arxiv.org/abs/1612.08220
3. S. M. Lundberg, S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 2017. arxiv.org/abs/1705.07874
4. L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, Annals of Mathematics Studies 28, pp. 307–317, Princeton University Press, 1953. doi.org/10.1515/9781400881970-018
5. J. Castro, D. Gómez, J. Tejada. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research* 36(5):1726–1730, 2009. doi.org/10.1016/j.cor.2008.04.004
6. E. Štrumbelj, I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 41(3):647–665, 2014. doi.org/10.1007/s10115-013-0679-x
7. R. Mitchell, J. Cooper, E. Frank, G. Holmes. Sampling permutations for Shapley value estimation. *Journal of Machine Learning Research* 23(43):1–46, 2022. arxiv.org/abs/2104.12199
8. M. D. Zeiler, R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2014. arxiv.org/abs/1311.2901
9. M. T. Ribeiro, S. Singh, C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of KDD*, 2016. arxiv.org/abs/1602.04938
10. I. Covert, S. Lundberg, S.-I. Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research* 22(209):1–90, 2021. arxiv.org/abs/2011.14878
11. J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of ACL*, 2020. arxiv.org/abs/1911.03429
12. W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, Z. Ren. Is ChatGPT good at search? Investigating large language models as re-ranking agents. In *Proceedings of EMNLP*, 2023. arxiv.org/abs/2304.09542
13. N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the ACL* 12:157–173, 2024. arxiv.org/abs/2307.03172