UNDERSTANDING RETRIEVAL AND REASONING IN LLM RERANKING UNDER AVAILABILITY BIAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used to rerank candidate items in recommender systems, where they may rely on both reasoning and retrieved evidence. We show that retrieval is not merely auxiliary to reasoning. Across two recommendation benchmarks, a small set of selected reviews matches a long reasoning trace on HR@1 at substantially lower inference cost, and the two signals combine. However, evaluating retrieval in this setting is confounded by a previously underappreciated bias. LLM rerankers systematically shift probability toward candidates that are accompanied by evidence-like text, even when that text belongs to another candidate or carries no useful information. We call this *availability bias*, by analogy to the availability heuristic in human judgment, without assuming a shared mechanism. Matching only the total number of retrieved passages does not remove the bias: when evidence is allocated unevenly, the allocation itself changes the ranking. We therefore introduce matched controls that separately account for evidence presence, count, allocation, content, and attribution. Under these controls, genuine evidence selection improves HR@1 by 0.04–0.10, while much of the apparent advantage among common retrieval heuristics is explained by where they allocate evidence rather than by which passages they select. Our results show that retrieval is a first-class component of LLM reranking, but also a potential source of hidden ranking manipulation. Reliable evaluation must control for evidence availability and attribution, or ranking gains may be misattributed to retrieved information rather than to how evidence is presented.

1 INTRODUCTION

Large language models are increasingly used as rerankers in search and recommendation systems. Initially, a candidate generation system proposes a small set of plausible candidates, and an LLM then uses the user's request, history, and candidate information to produce a final ranking (Hou et al., 2024b; Sun et al., 2023). At this stage, performance can be improved through at least two distinct mechanisms. The model can spend additional inference compute on reasoning (Weller et al., 2025; Zhuang et al., 2026), or it can be given additional evidence retrieved for the candidates, such as reviews, descriptions, or passages (Kim et al., 2025; Wei et al., 2025; Hwang et al., 2026). These mechanisms are usually studied separately, yet both change the information available to the final decision. In recommendation in particular, retrieved evidence is commonly attached to individual candidates, each candidate's block drawn from its own reviews or metadata (Wei et al., 2025; Hwang et al., 2026), so retrieval determines not only what the model reads but also which candidate each piece of evidence is presented as supporting.

This coupling creates a basic identification problem when evaluating retrieval. A comparison between a reranker with retrieved evidence and the same reranker without it attributes the entire change in ranking quality to the information contained in the retrieved text. We show that this interpretation can be wrong: merely placing evidence-like text under a candidate systematically shifts probability toward that candidate, even when the text describes another candidate or carries little useful information. We call this phenomenon *availability bias*, borrowing the term from the availability heuristic of human judgment while making no claim of a shared mechanism. (Tversky & Kahneman, 1973; 1974) Consequently, matching only the total amount of retrieved text is not sufficient to isolate retrieval quality. When evidence is distributed unevenly across candidates, its allocation can

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

itself change the ranking. Evaluating retrieval therefore requires controls that separate the effects of evidence presence, allocation, content, and attribution.

We measure what that difference is made of, on a task curated for this purpose, and report four findings, one per section.

1. **Retrieval is helpful, and competitive with reasoning at a fraction of the cost** (§3, Figure 1). Ten reviews chosen by a selection rule, read with no reasoning, tie a 2,800-token reasoning trace on HR@1 in every setting and lead on NDCG@10, in a quarter of a second against a minute; the two levers stack; retrieving after reasoning, for a shortlist, is never ahead of retrieving before the read.
2. **The reader has availability bias** (§4, Figure 2). Under the three candidates it cannot tell apart, a candidate's own review and another candidate's review move the odds the same way, a content-free block of matched length moves them a third as far, a fabricated review makes an arbitrary candidate the top choice three times in four, and passages retrieved for the request but filed under the wrong candidates are worse than none. The reader credits a candidate with whatever is written under it, and the controls that measure this are what let evidence be credited correctly.
3. **Matching the amount of text is not enough** (§5). Arbitrary reviews at the same count as a rule are no better than none, worse at small budgets, and cost more; the same holds on a shortlist.
4. **Selection is where the gain is** (§6). Rules that hold the count per candidate equal and rules that spend one shared budget unevenly are worth $+0.04$ to $+0.10$ HR@1 over the count-matched control; within the family we try the objective is second-order, and an oracle is a further 0.03 to 0.07 above the best rule. The rules and what inspired them are in Appendix C.

All of this is measured on a task we curated so that evidence is the deciding signal (§2, Appendix A). An instance is a user, a request, and twenty candidates, of which one is the item the user chose. The candidates mimic what a classic recommender's first stage hands to a reranker: the nineteen negatives are items an item–item embedding model ranks near the target, drawn from a fixed window of that ranking so that they are neither trivially different from the target nor near-duplicates of it, and matched to the target in popularity so that popularity alone scores at chance. The request is a paraphrase of the user's own text, written by a different model family from the reader, and every candidate carries a pool of up to twenty reviews from which evidence is drawn.

Our contributions are the demonstration of availability bias and the family of controls that measures it; the count-matched protocol, released as a tested implementation, under which a selection rule is credited only for what its choices inform; the curated task and its audits; the measurement under the protocol of eight selection rules and of selected evidence against reasoning; and the negative result for retrieval after reasoning.

2 A TASK THAT MIMICS A RECOMMENDER'S SHORTLIST

The task. An *instance* is one user, one request, and twenty candidate items, of which exactly one is the item the user chose. The reader sees the user's recent history, the request, and the twenty titles in a shuffled order with two-digit labels, and is asked for a ranking. We read the ranking out pointwise: the *score* of candidate i is the model's log-probability of the answer `Ranking: i`, computed for all twenty in one forward pass, so the reader can never name an item off the list and never abstains (Hou et al., 2024b). Decoding is greedy throughout. The reader is Qwen3-8B (Qwen Team, 2025) unless stated. A *direct read* answers at once: the forced readout with no thinking block. A *reasoning read* opens the model's thinking block first and reads the same answer out after it.

Two corpora. **Amazon-20** is built from Amazon reviews in three categories (Toys and Games, Grocery, Office Products); **Reddit** from the Reddit-Movie recommendation threads of He et al. (2023), whose requests are written by people asking for suggestions. Each holds 750 instances. A user with a sufficient history contributes one: the item they chose next is the target and their previous items are the history. The nineteen negatives are what a classic recommender's first stage would put beside the target. They are mined from the same category by an item–item embedding

model (Hou et al., 2024a), from a fixed window of its similarity ranking (ranks 200 to 400), so that they are neither trivially different from the target nor near-duplicates of it, and they are matched to the target in popularity: the number of negatives more popular than the target is drawn uniformly, so that the target's popularity rank is uniform over the twenty and a ranker that orders by popularity alone scores at chance (0.068 and 0.044 HR@1 against 0.341 before matching on Amazon). The request is a paraphrase of the user's own text, the target's review on Amazon and the opening post on Reddit, written by Llama-3.1-8B with a mechanical check that it does not name the item (nor its store, director or cast), so that it reads like a user's request and not like the reader's own distribution. Every candidate carries a *pool* of up to twenty reviews. Appendix A gives the construction in full with its audits.

Metadata. Items also carry structured metadata: price, store and brand on Amazon; year, director and cast on Reddit. It is factual and standardised, the same fields for every item, so we treat it as a property of the prompt rather than as something to retrieve: a setting with titles only and a setting that adds the metadata of every candidate. It costs no selection, and a reasoning read can check it against a title, which is why, if any input favours reasoning over retrieval, it is this one; every experiment reports both settings. The retrievable evidence is the review pool, which is per item, free text and unequal, and is what a selection rule chooses from.

Evidence. A *block* is the text attached under one candidate: one or more of its reviews, formatted with the candidate's label. The *shown set* is the set of candidates that receive a block. A *budget* B is the total number of passages the prompt carries. The *allocation* is how those B passages are spread over candidates: *one each*, so that the shown set is the whole shortlist and every candidate carries exactly one, or a *shared budget*, so that a candidate may carry several or none and the shown set is whatever the rule produces. *Coverage* is uniform when every candidate carries the same number of passages.

Two designs. Evidence is retrieved *before the read*, in the one read over all twenty candidates, or *after reasoning*: the reader first reasons over the twenty without evidence, its top K form a *shortlist*, evidence is retrieved for the shortlist, and a second read re-ranks it, a candidate off the shortlist keeping its first-stage rank. Every configuration is named by its two levers, whether it retrieves and whether it reasons. §§5–6 measure selection rules on the reader's own top K, because that is where the literature attaches evidence, and on the full list of twenty.

Statistics. We report HR@1 over the twenty candidates, and HR@10, MRR and NDCG@10 where they add something. Every contrast is *paired*: two configurations see the same instances, the same shuffled presentation and, in the pipeline, the same shortlist, so a contrast is a within-instance difference with a 95% percentile-bootstrap interval over instances (20,000 resamples). Rules run together as one comparison form a *family* and are corrected with Holm's procedure within it; in tables, a bold entry survives the correction, and an italic one has an interval that excludes zero without surviving. Byte-identical prompts read twice disagree on about 10% of instances because of nondeterministic kernels, which puts a *floor* of roughly 0.03 on an HR@1 difference we will call real at $n \approx 700$; a point estimate inside the floor is reported as inside the floor whatever its interval.

3 Finding 1: retrieval is helpful, and competitive with reasoning at a fraction of the cost

A reranker can pull two levers, retrieved evidence and a reasoning trace, and Figure 1 puts every combination of them on the same instances against what each costs. Retrieval alone stands level with reasoning alone, at a quarter of a second against a minute: paired on the same instances, the HR@1 difference between the rows "retrieval, no reasoning" and "reasoning, no retrieval" of Table 1 is inside the readout floor in all four settings, and on NDCG@10 retrieval is ahead in all four. The two levers stack: reasoning over the retrieved reviews is the best configuration in every setting, by $+0.03$ to $+0.05$ HR@1 over retrieval alone, and its trace is shorter than a trace over titles alone.

Two things might have favoured reasoning and do not. Metadata should: a trace can check a year or a price against a title, yet the tie holds in the two "+ metadata" blocks of Table 1. The 4,096-token cap on the trace does not bind: raising it to 8,192 leaves HR@1 within 0.001 (Appendix E).

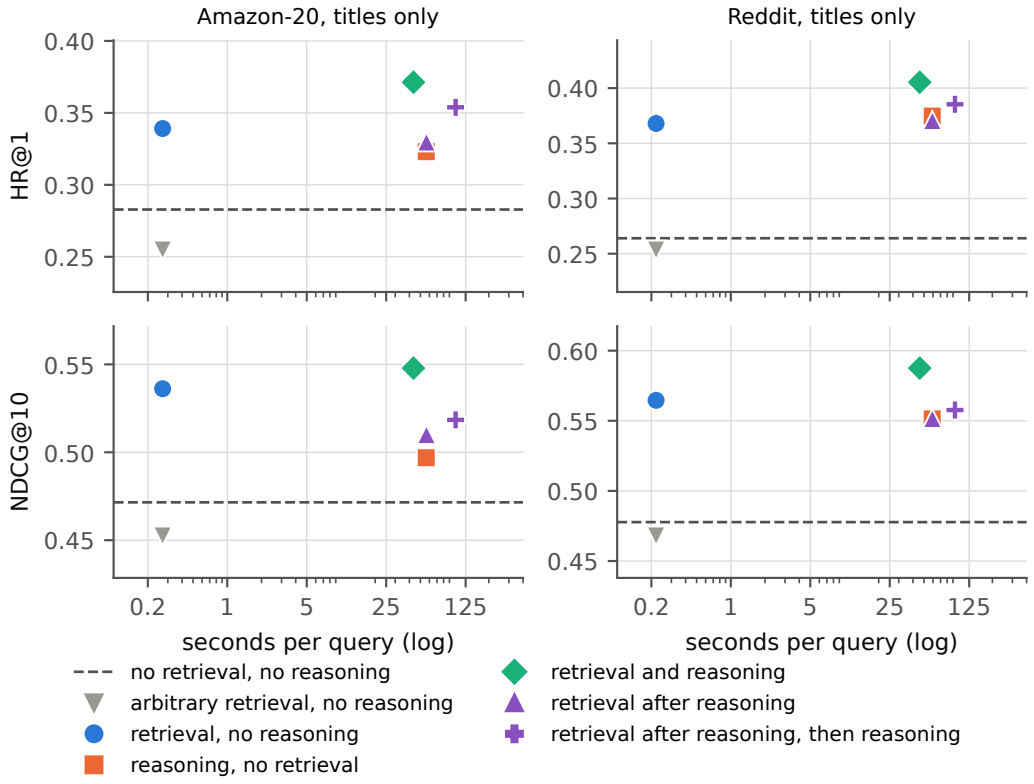

Figure 1: **Ranking quality against cost for every configuration of the two levers.** HR@1 (top) and NDCG@10 (bottom) over twenty candidates against seconds per query on one idle H100 (medians over twenty timed queries); Qwen3-8B, titles only, n=746 and 750. The dashed line is the reader with neither lever. Retrieval is ten reviews chosen by a selection rule; arbitrary retrieval is ten drawn at random, plotted at the retrieved read's cost, whose prompt size it matches; reasoning is a thinking trace before the same readout; retrieval after reasoning is the pipeline that reasons, keeps its top ten, retrieves for them and reads them again. Paired intervals are in Table 12.

Neither reader size nor the order of the levers changes the picture. Retrieval helps at every size and most for the smallest, while reasoning alone grows with capacity, so selected reviews are the stronger input at every size we ran. Reasoning first and then retrieving for the shortlist buys nothing that retrieving before the read does not: it is behind or level in all eight paired comparisons and, when both reads reason, costs about twice the tokens (Appendix E, Table 10).

The comparison presupposes the selection. The row "arbitrary retrieval, no reasoning" is below the reader with neither lever in every setting on a direct read, and in three of four on a reasoning read. Retrieval is a lever only when what is retrieved is selected and correctly attributed, and the next three sections are about that qualification.

4 Finding 2: Availability Bias, and the Controls That Measure It

Before asking which evidence helps, we ask what attaching evidence does on its own. The shown set is the three candidates the reader's own reasoning ranks highest: the ones it cannot tell apart, which together hold 0.72 of its probability on a direct read. Each receives a block, appended to the prompt after the candidate list under a heading, and the twenty-candidate prompt is read again in a fresh direct session, so no trace ever sees the evidence and the reasoning read serves only to choose the three. We measure how far the odds that the answer is one of the three move.

Table 1: **Every configuration of the two levers: six metrics and cost.** Qwen3-8B, twenty candidates, a budget of ten passages chosen by the contrastive rule (retrieval) or drawn at random (arbitrary retrieval). Tokens are per query; seconds are medians over twenty timed queries on an idle H100, titles only.

configuration	HR@1	HR@10	MRR	NDCG@10	prompt	generated	s/query
Amazon-20, titles only ($n = 746$)							
no retrieval, no reasoning	0.283	0.718	0.416	0.472	1,370	0	0.21
arbitrary retrieval, no reasoning	0.255	0.717	0.391	0.452	2,733	0	–
retrieval, no reasoning	0.339	0.779	0.476	0.536	2,836	0	0.27
reasoning, no retrieval	0.323	0.724	0.447	0.497	1,370	2,881	56.27
retrieval and reasoning	0.371	0.775	0.494	0.548	2,836	2,717	43.47
Amazon-20, titles + metadata ($n = 746$)							
no retrieval, no reasoning	0.264	0.724	0.408	0.468	5,963	0	–
arbitrary retrieval, no reasoning	0.253	0.717	0.397	0.457	7,333	0	–
retrieval, no reasoning	0.328	0.768	0.465	0.524	7,429	0	–
reasoning, no retrieval	0.308	0.756	0.445	0.506	5,963	3,326	–
retrieval and reasoning	0.382	0.816	0.513	0.575	7,429	2,868	–
Reddit, titles only ($n = 750$)							
no retrieval, no reasoning	0.264	0.748	0.413	0.478	554	0	0.19
arbitrary retrieval, no reasoning	0.253	0.728	0.406	0.468	1,288	0	–
retrieval, no reasoning	0.368	0.789	0.509	0.565	1,384	0	0.22
reasoning, no retrieval	0.375	0.764	0.502	0.551	554	2,780	59.12
retrieval and reasoning	0.405	0.803	0.534	0.588	1,384	2,569	45.89
Reddit, titles + metadata ($n = 750$)							
no retrieval, no reasoning	0.276	0.696	0.405	0.456	2,453	0	–
arbitrary retrieval, no reasoning	0.261	0.708	0.406	0.462	3,185	0	–
retrieval, no reasoning	0.367	0.788	0.498	0.555	3,284	0	–
reasoning, no retrieval	0.357	0.757	0.480	0.533	2,453	2,993	–
retrieval and reasoning	0.405	0.787	0.528	0.578	3,284	2,723	–

Four contents are compared (Appendix B shows each on an instance). *Real* is the candidate's own most-voted review, or its metadata. *Swapped* keeps the three headings and gives each the content of the next shown candidate, so every block sits under an item it does not describe while the prompt contains the same text. *Filler* keeps the heading and replaces the content with neutral boilerplate of matched word count. *Template only* keeps the heading and a line saying there is nothing. Figure 2 is the result.

Wrong attribution. The swapped block moves the odds as far as the real one, for reviews and for metadata, on both corpora; for reviews the paired difference has an interval containing zero on both (Table 7). Another item's price and store serve a candidate as well as its own. The one exception is film metadata on Reddit: a year, a director and a cast can be compared with a title, and there the swapped block moves the odds less than the real one. A product's price cannot be checked, and a review cannot be checked anywhere. The reader verifies where verification is possible, and nowhere else.

Even nothing. Filler of matched length, which says nothing about any product, moves the odds a third as far as a real review (0.29 and 0.36 of the real shift on the two corpora for reviews, 0.31 and 0.22 for metadata); only the bare heading, with a line saying there is nothing, moves them back.

The effect is not unique to this setup; under a second reader from another family, Llama-3.1-8B-Instruct, shows every sign of it with larger sizes: under its own bottom three, another item's metadata moves so much probability that its HR@1 over the twenty drops from 0.264 to 0.169. At the extreme, a single review fabricated to claim that an arbitrary non-target satisfies the request makes it the first choice in 0.749 of instances, against 0.286 before any evidence (Appendix D).

The blocks above seldom bear on the request, and the reader credits them anyway. When the passages are the ones a selection rule retrieved *for the request*, the same credulity has a cost: on a read of all twenty candidates, the rule's ten passages raise HR@1 by $+0.06$ and $+0.10$ on the two corpora,

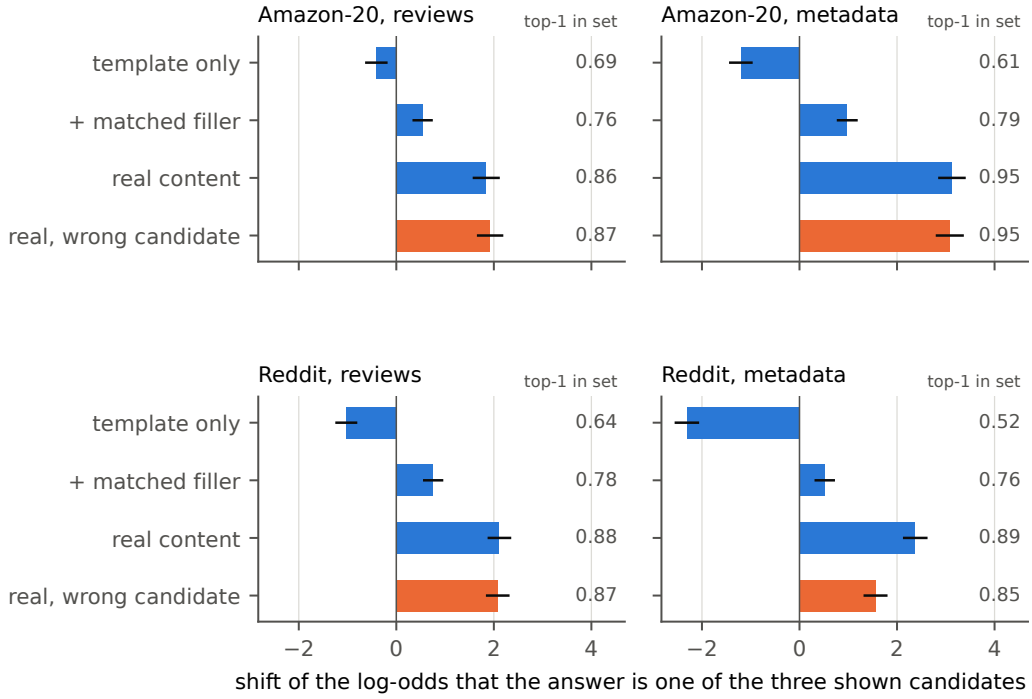

Figure 2: **Under the three candidates the reader cannot tell apart, evidence filed under the wrong one moves the odds as far as correct evidence, and a block of nothing moves them a third as far.** The three candidates the reader's own reasoning ranks highest each receive a block whose content varies. The axis is the change in the log-odds that the answer is one of the three, evidence against none, on a fresh direct read, with 95% intervals over instances; the right-hand column is how often the top-ranked candidate is one of the three afterwards. Qwen3-8B

and the same ten re-filed among the candidates they covered are no better than ten arbitrary passages, on Reddit worse than none (Appendix E). A review that says nothing is credited to its candidate; a review that says something is credited to whichever candidate it sits under. So a comparison of "with evidence" against "without" credits the evidence for two things, what the text told the reader and the fact that some candidates carried text, and a control must hold the second fixed for the first to be measured.

The controls. A selection rule is compared against a prompt that carries the *same amount of text, placed without reference to the request*, and the control holds fixed everything the rule was not allowed to decide. Table 2 lists them. *Presence-matched* puts one arbitrary passage under every shortlisted candidate, the counterpart of a rule that gives every candidate one passage. *Count-matched* scatters B arbitrary passages over the instance, the counterpart of a rule that spends a shared budget. *Content-resampled* keeps the rule's own allocation and refills every slot at random from the same candidate's pool, so that a shared-budget rule's gain splits into the part that came from where the budget went and the part that came from the text it chose. *Swapped* is the control used above and measures attribution. Every control is drawn from a seed fixed per instance; Appendix B shows each on an instance.

5 Finding 3: Matching the amount of text is not enough

If attached text moves the reader on its own, the natural repair is to give every configuration the same amount of it, and credit a rule only against a prompt with the same number of arbitrary passages. That is the right control, and the rest of the paper uses it. It is not a method.

Table 2: **The controls.** Each holds fixed one more thing than the last; a rule is credited with the contrast against the control that matches what it was allowed to decide.

control	holds fixed	varies	*rule* − *control* measures
`none`	—	everything	presence and selection together; never the contrast
presence-matched	count and coverage: one passage per candidate	which passage	a one-each rule's choice of passage
count-matched	the total count	which passages, and under which candidates	a shared-budget rule's choices, both kinds
content-resampled	the rule's own allocation	the text in each slot	the part of a shared-budget rule's gain that is content
swapped	the rule's own passages	their attribution	what correct attribution is worth

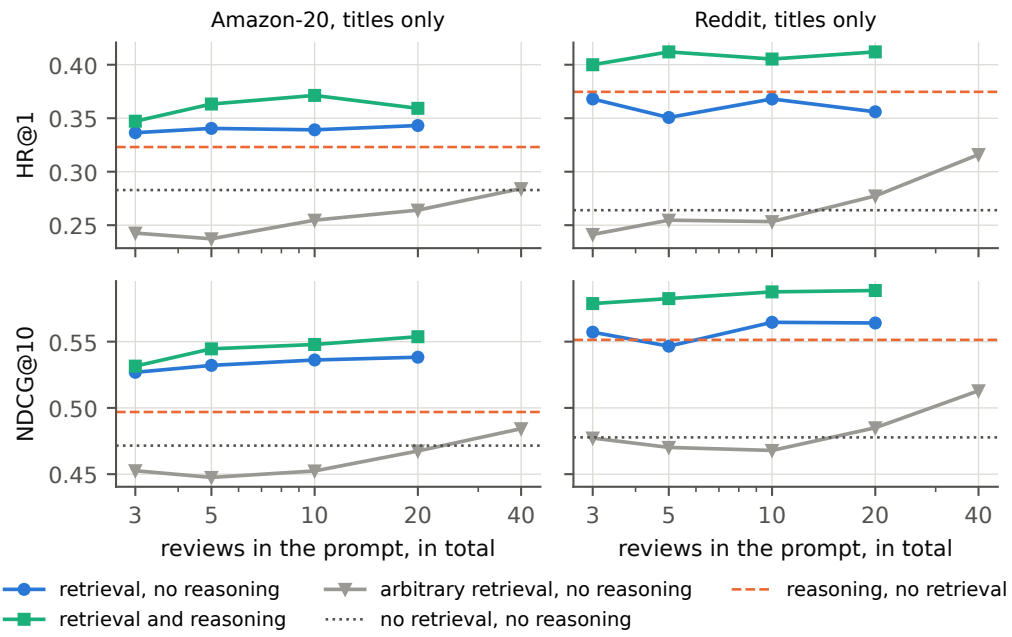

Figure 3: **Arbitrary against selected evidence at every budget.** HR@1 (top) and NDCG@10 (bottom) over twenty candidates against the total number of reviews in the prompt; Qwen3-8B, titles only, n=746 and 750. The dotted and dashed lines are the reader with neither lever and with reasoning alone. Arbitrary retrieval (grey) is at or below no retrieval up to twenty reviews and above it only at forty; ten selected reviews (blue) beat forty arbitrary ones.

One arbitrary passage under every candidate is worth nothing: the presence-matched control minus `none` has an interval containing zero in all eight shortlist cells we run (Table 6) and on the full list of twenty. A shared budget of arbitrary passages is worse than nothing until it approaches uniform coverage: at three, five and ten passages over twenty candidates it is below no evidence in every setting, surviving correction at the small budgets, and it reaches the no-evidence level only at forty, two per candidate on average (Figure 3); on the reasoning read's shortlist of ten, ten arbitrary passages are within 0.02 of none (the paired contrasts are in Appendix E). The shape is availability bias at work: a scarce budget placed at random lands under a random subset and moves the odds towards it on no grounds, and a budget that covers everyone has nothing to favour. The rise at forty may also be an artefact of pool size: with about twenty reviews per candidate, forty draws over twenty candidates most probably include the relevant review already, so past a point the arbitrary curve is no longer arbitrary.

Ten arbitrary reviews add 732 to 1,370 prompt tokens per query and forty add 1,798 to 2,916, in the four settings (Table 1, column "prompt"); ten reviews chosen by a selection rule cost what ten arbitrary ones cost and beat forty arbitrary ones in every setting. On one idle H100, ten arbitrary reviews make the direct read a third slower than no evidence and forty nearly twice as slow (0.29 and 0.41 seconds per query against 0.21 on Amazon-20; 0.24 and 0.32 against 0.19 on Reddit).

6 Finding 4: selection is where the gain is

Measured against the control that matches what it was allowed to decide, does a selection rule buy anything? We ran eight: four objectives, relevance, contrastive relevance, maximal marginal relevance and a determinantal rule, each scoring a passage against the request through one small sentence encoder, under each of the two allocations of §2: one passage per candidate, credited against the presence-matched control, or a shared budget of K passages the rule may spend unevenly, credited against the count-matched control. Appendix C specifies each rule and the work that inspired it.

Selection is worth $+0.04$ to $+0.10$ HR@1 on the reader's own shortlist where the count is worth nothing: every rule with a shared budget survives correction in several of the eight cells, and with one passage per candidate the two diversity objectives survive on Reddit at $K{=}10$. On the full list of twenty, where coverage is uniform and there is nothing for availability to favour, one passage per candidate chosen by relevance is worth $+0.06$ and $+0.05$ over the presence-matched control on the two corpora, and a shared budget of ten is worth as much as a passage for each; three presentation seeds reproduce the shortlist effects (Appendix F).

Within the family. Where the budget goes matters more than how a passage is scored. Moving a rule from one passage each to a shared budget is worth about five times what changing its objective is; keeping a shared-budget rule's allocation and refilling every slot from the same candidate's pool costs nothing, while the allocation part of the same gain survives correction in eleven of sixteen combinations; and allocations designed from the reader's own first-stage scores, with no encoder in the split, reproduce the gain, while starving the candidate the reader already prefers is harmful in every cell (Appendix C). This is availability bias working for the rule: a shared-budget rule looks informative when it concentrates evidence under the candidates the reader already leans to.

The room above. An oracle that may look at the answer leaves 0.03 to 0.07 HR@1 above the best rule at $K{=}10$, and every selector we tried beyond the family, a cross-encoder, an in-domain contrastive encoder, a set-level score, and the reader itself asked to choose, either failed to beat cosine at an upstream gate or tied it in front of the reader (Appendix C); none was trained for this task. A selector that does better would have to change *which candidates* carry evidence in a way that tracks the target rather than the reader's prior, or change *what* is shown in a way the reader uses when coverage is uniform. The controls of §4 measure both, and they are the terms on which such a selector should be reported. Table 6 in Appendix C gives every rule in every cell, with the oracle and the presence term beneath.

7 Related work

Judgment under availability. Tversky & Kahneman (1973) named the *availability heuristic*: people judge the frequency of a class or the probability of an event by the ease with which instances come to mind, so a class whose instances are easier to retrieve is judged more numerous (Tversky & Kahneman, 1974). Tversky & Kahneman (1983) showed the related conjunction fallacy: a description which fits raises the judged probability of an option even where it cannot raise its actual probability, committed in the Linda problem by 85% of respondents in a direct test and by statistically trained respondents as often as by naive ones. We borrow the name and not the mechanism: what we measure is that evidence-like text present under a candidate moves the reader's probability onto it irrespective of whether it is about that candidate, a bias of availability in the plain sense of the word.

Language models as rerankers. Listwise prompting with a permutation read out of the generation was established by Sun et al. (2023) and applied to recommendation by Hou et al. (2024b), who

documented that free-form generation names items off the candidate list on about 3% of queries, which is why we read the ranking out pointwise. Pairwise (Qin et al., 2024) and setwise (Zhuang et al., 2024) prompting trade forward passes for calibration; we hold the readout fixed so that only the prompt's content varies. Reasoning rankers (Weller et al., 2025; Zhuang et al., 2026) find their gains concentrated on reasoning-heavy retrieval rather than classical relevance, consistent with the stacking we see: the trace is worth most when the prompt is thin. Recent rerankers attach review- or metadata-derived text under each candidate (Kim et al., 2025; Wei et al., 2025; Hwang et al., 2026); Hwang et al. (2026) report that appending all of an item's metadata falls below titles alone, an effect our count-matched control isolates.

Reviews as evidence. Review text has been used to predict ratings since Zheng et al. (2017) and Catherine & Cohen (2017); Sachdeva & McAuley (2020) showed that for models with bias terms, replacing the reviews with empty strings barely changes accuracy, an early warning that a gain attributed to text may not come from its content. Ni et al. (2019) study which parts of a review justify a recommendation. Our controls are that concern turned into a protocol for the reader rather than the model: the same amount of text, chosen without regard to the request.

Noise, position and distraction. Cuconasu et al. (2024) found that adding random passages to a retrieved context can raise a reader's accuracy by 0.07 absolute for Llama2-7B beside four retrieved documents, in a setting where more than half the queries have no answer-bearing document, and that with a gold passage present the effect depends on the model and on where the random passages sit. Liu et al. (2024) showed accuracy falling for evidence placed mid-context and Shi et al. (2023) showed that a single irrelevant sentence degrades arithmetic reasoning. The bias we measure is the ranking counterpart of these: the effect of attached text that is not about the candidate, isolated from the effect of text that is.

Selection rules. Maximal marginal relevance (Carbonell & Goldstein, 1998), graph centrality for generic extractive summarization (Erkan & Radev, 2004) and determinantal point processes (Kulesza & Taskar, 2012; Chen et al., 2018) are the objectives our rules instantiate (Appendix C). Our contribution is not a rule but the measurement: under a matched control, selection is worth several times the count, and the rules in common use separate on their allocation rather than their objective.

8 Limitations

The availability experiment is measured on one reader family with a direct readout in the main text; a second reader is in Appendix D, and the size of the effect differs between corpora and evidence types even where its sign does not. Retrieval here selects within a candidate's own pool of reviews, not from an open corpus, and twenty candidates is a short list; a list long enough that evidence for every candidate does not fit one prompt is where retrieval after reasoning has its best argument and where we have no measurement. The selection finding is scoped as stated in §6: one encoder, relevance-shaped objectives, the reader's own shortlist, and no learned selector. The requests of both corpora were paraphrased by a Llama-3.1-8B model during construction, so the Llama reader we report in the appendix is not independent of either corpus.

9 Conclusion

Retrieved evidence is worth as much to a language-model reranker as a reasoning trace, at a fraction of a percent of its latency, and the two stack. But the reader believes what is put under a candidate: attached text moves probability onto it whether or not it is about it, so retrieving everything promotes on no grounds, and matching the amount of text does not repair it, because arbitrary text at a matched count is worth nothing and costs. What is worth $+0.04$ to $+0.10$ HR@1 over that control is selection, and the room above the rules we tried is still large. Mind what your model reads; the controls here are how to tell whether it read anything at all.

AI USE STATEMENT

Generative AI tools were used to write and refactor experiment code and the scripts that generate tables and figures from the result files, and for wr. All AI-assisted code was reviewed and tested by the authors; every number in the paper is produced from a result file by a script and checked against that file automatically; all claims and the final text were reviewed by the authors, who take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix A gives the construction of both corpora with its audits; Appendix B shows every control on an instance; Appendix C specifies every selection rule; Appendix G describes the released implementation of the controls, the paired contrast and the family correction. The task data, the result files and the scripts that generate every table and figure are released with the code.

REFERENCES

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A human generated MAchine Reading COmprehension dataset. *arXiv preprint arXiv:1611.09268v3*, 2018. URL https://arxiv.org/abs/1611.09268v3.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pp. 335–336, Melbourne, Australia, 1998. ACM. doi: 10.1145/290941.291025.

Rose Catherine and William Cohen. TransNets: Learning to transform for recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*, pp. 288–296, Como, Italy, 2017. ACM. doi: 10.1145/3109859.3109878.

Laming Chen, Guoxin Zhang, and Hanning Zhou. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/dbbf603ff0e99629dda5d75b6f75f966-Abstract.html.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, pp. 719–729, Washington, DC, USA, 2024. ACM. doi: 10.1145/3626772.3657834.

Güneş Erkan and Dragomir R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004. doi: 10.1613/jair.1523.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, pp. 720–730, Birmingham, United Kingdom, 2023. ACM. doi: 10.1145/3583780.3614949.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952v1*, 2024a. URL https://arxiv.org/abs/2403.03952v1.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (eds.), *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II*, volume 14609 of *Lecture Notes in Computer Science*, pp. 364–381. Springer, 2024b. doi: 10.1007/978-3-031-56060-6_24.

Dojun Hwang, Seunghan Lee, Cheonyoung Park, Sara Yu, and SeongKu Kang. Profiling what matters: Context-aware item profiles from large-scale metadata for LLM recommenders. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26)*, 2026. doi: 10.1145/3799682.3840943.

Jieyong Kim, Hyunseo Kim, Hyunjin Cho, SeongKu Kang, Buru Chang, Jinyoung Yeo, and Dongha Lee. Review-driven personalized preference reasoning with large language models for recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, pp. 1697–1706, 2025. doi: 10.1145/3726302.3730055.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1018. URL https://aclanthology.org/D19-1018/.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1504–1518, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.97. URL https://aclanthology.org/2024.findings-naacl.97/.

Qwen Team. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Noveen Sachdeva and Julian McAuley. How useful are reviews for recommendation? a critical review and potential improvements. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, pp. 1845–1848. ACM, 2020. doi: 10.1145/3397271.3401281.

Sentence-Transformers. all-MiniLM-L6-v2. Hugging Face model card, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2021.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and

Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 14918–14937, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.923. URL https://aclanthology.org/2023.emnlp-main.923/.

Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315, 1983.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Kaiwen Wei, Jinpeng Gao, Jiang Zhong, Yuming Yang, Fengmao Lv, and Zhenyang Li. Learning to shop like humans: A review-driven retrieval-augmented recommendation framework with LLMs. *arXiv preprint arXiv:2509.00698*, 2025.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. Rank1: Test-time compute for reranking in information retrieval. In *Proceedings of the Second Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=Pg0PAvbhGv.

Lei Zheng, Vahid Noroozi, and Philip S. Yu. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*, pp. 425–434, Cambridge, United Kingdom, 2017. ACM. doi: 10.1145/3018661.3018665.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, pp. 38–47, Washington, DC, USA, 2024. ACM. doi: 10.1145/3626772.3657813.

Shengyao Zhuang, Xueguang Ma, Zheng Yao, Shuai Wang, Bevan Koopman, Jimmy Lin, and Guido Zuccon. Rank-R1: Enhancing reasoning in LLM-based document rerankers via reinforcement learning. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*, pp. 4419–4425, Melbourne, VIC, Australia, July 2026. ACM. doi: 10.1145/3805712.3809961.

A THE TASK IN FULL

The construction is the same on both corpora; the two differ in their sources and in the popularity field. Table 3 gives the resulting task in numbers. Every step is a script in the released repository, seeded so that the benchmark is reproducible row for row.

Table 3: **The task in numbers.** Sizes, evidence depth and the shortcut audits. A shortcut ranker sees no evidence: popularity orders the twenty by the popularity field; the constant-slot policy always picks the presentation slot that is most often the target, which is the null a label-order bias would reach; lexical overlap orders by word overlap between the request and the title. Chance is 0.050.

	Amazon-20	Reddit
Source		
corpus	Amazon Reviews 2023, three categories	Reddit-Movie, five subreddits
candidate miner	BLaIR item–item cosine, ranks 200–400	BLaIR over film metadata, ranks 200–400
popularity rank of the target	uniform over 1–20	uniform over 1–20
popularity field matched	rating count	mention count
request	target review, paraphrased	opening post, paraphrased
Size		
instances evaluated	750	750
candidates per instance	20	20
reviews per candidate, mean (median)	15.25 (20)	16.08 (20)
history length, median	14.5	12
request length in words, median	23	38
Shortcut audits (HR@1 over twenty; chance 0.050)		
popularity alone, before matching	0.341	–
popularity alone, after matching	0.068	0.044
best constant-slot policy	0.072	0.064
lexical overlap with the request	0.152	0.059
mean cosine of negatives to the target	0.894	0.877

The candidates. A classic recommender's first stage returns items that are close to what the user is likely to want, and a reranker's job is to order those. We therefore mine the nineteen negatives of an instance from an item–item similarity ranking rather than at random: an embedding model trained on review–metadata pairs, BLaIR (Hou et al., 2024a), encodes every item's text (CLS-pooled, L_2-normalised), and the negatives are drawn from ranks 200 to 400 of the target's cosine-sorted same-category neighbours. The nearest neighbours are excluded because they are near-duplicates (variants, bundles, the same film under another title) that no request could separate; far items are excluded because a title alone separates them. The window doubles when the popularity match below cannot be filled inside it. On Amazon this places the negatives at a mean cosine of 0.894 to the target; on Reddit, whose item pool is the 5,296 films with a usable comment pool, rank 200 is already the fourth percentile of the neighbourhood, and the mean cosine is 0.877.

Popularity. Ordering the twenty by popularity alone reaches 0.341 HR@1 on Amazon before matching (Table 3): the target is usually among the better-known items in its neighbourhood, so a ranker that reads nothing would score above every evidence rule we run. The negatives are therefore matched to the target in popularity, on $\log_{10}(1 + \text{rating count})$ for Amazon and $\log_{10}(1 + \text{mention count})$ for Reddit. The number of negatives more popular than the target is drawn uniformly from 0 to 19, so that the target's popularity rank is uniform over the twenty; the negatives on each side are drawn at random from the three times as many window items closest to the target in log-popularity, and any side that cannot be filled is completed from the other, closest first. A fixed half-above, half-below split would instead make popularity *anti*-informative, since the target could then never be the most popular candidate. After matching, popularity alone scores 0.068 on Amazon and 0.044 on Reddit, against a chance rate of 0.050 and a constant-slot policy of 0.072 and 0.064.

Users and targets. The source is the Amazon Reviews 2023 release, the corpus BLaIR was trained on (Hou et al., 2024a), in its five-core, leave-one-out form, in three categories: Toys and Games, Grocery and Gourmet Food, Office Products. Per category, 1,500 users with at least ten interactions are sampled with a fixed seed; the user's last interaction in time order is the target and the twenty most recent before it are the history. The candidate pool is every item with at least five ratings plus any item a sampled user touched. Of 506, 426 and 482 instances built, the first 250 per category are evaluated.

The request. The request is written from the target's own review, which must have at least 30 words, by Llama-3.1-8B-Instruct at temperature 0.7 and top-p 0.9 (at most 90 new tokens; the review is truncated to 1,500 characters). A generated request that contains the store's name, or any title token longer than seven characters or containing a digit, is regenerated up to twice and the instance is dropped if a leak-free request is not obtained. The rephraser is deliberately a different model family from the reader: a request phrased in the reader's own distribution would flatter it.

The review pools. A candidate's pool holds its verified-purchase reviews ranked by helpful votes (ties broken by a seeded key), each truncated to 400 characters. The benchmark was first collected with five reviews per item; the pools used throughout this paper hold twenty, built by re-streaming the corpus with a check that the published five are, row for row, the top five of the deeper pool. Of 15,000 candidates, 227 had no verified review at all and were given unverified ones so that every candidate can carry evidence; four instances still contain a candidate with an empty pool and are dropped from every Amazon comparison, which is why Amazon cells report n=746. A review is removed from a pool if it was written by the instance's own user or if its token set overlaps the source review at a Jaccard of 0.6 or more, so that the request's source can never be retrieved as evidence.

Reddit. The source is Reddit-Movie (He et al., 2023), 1,669,720 turns from five film-recommendation subreddits. Two tiers of turn are kept. A *source* turn names exactly one film, has at least 30 words and at least one sentence of 12, and carries at least two upvotes; it supplies the target, the history and the request. An *evidence* turn names exactly one film, has at least 8 words, is not deleted and is deduplicated; it supplies the comment pools, ranked by upvotes and capped at twenty per film. The item embedding for negative mining is BLaIR over each film's Wikidata metadata rather than over its comments, so that the negatives are not mined from the same text later shown as evidence. Popularity is the film's mention count. An instance is kept only if every candidate has at least five comments. The request is the paraphrase of the asker's opening post, by the same Llama-3.1-8B-Instruct, with the leak check extended to the film's director and cast. Three rules remove a comment from a pool: it is the source turn or a near-duplicate of it, it is by the same redditor, or it is from the same conversation. Of 3,792 instances built, 750 are sampled with a fixed seed with no user appearing twice.

Prompt settings. Titles only shows each candidate as its label and title. The metadata setting adds, under every candidate alike, the item's structured fields: on Amazon the price, average rating, rating count, store, categories, features and description; on Reddit the year, runtime, genres, director, cast and country. The user's history is shown as titles in both settings.

B WHAT EACH CONTROL LOOKS LIKE

Every control is a transformation of the evidence block, the part of the prompt that follows the candidate list. The listings below are generated from one Amazon-20 instance (Toys and Games, instance 0) by the same functions the experiments call; reviews are shortened for display and marked `[...]`, and the reader sees them in full, up to 400 characters each. The request and the candidate list are shown once; the evidence block follows them in every prompt.

The four contents (§4). Three candidates receive a block. *Real* is the candidate's own most-voted review, with its star rating. *Swapped* keeps the three headings and gives each the content of the next shown candidate in the set, cyclically, so the prompt contains exactly the same text and only the headings differ. *Template only* keeps the heading and replaces the content line with the phrase "(no information available)". *Matched filler* keeps the heading and replaces the content with the same number of words, drawn in order from one fixed passage of administrative boilerplate that says nothing about any product ("this entry was last checked by the catalogue team and the record below was filed for internal use only …"), repeated as needed; the word count is matched per block, so a long review gets a long filler. For metadata the same four transformations are applied to the item's structured fields.

The request and the candidate list (shortened)

```
What the customer is looking for:
I need sleeves for Pokemon cards that lay flat so my kid can easily put
    them in a binder.
```

```
Candidate products to rank (indices 00-19):
[00] Star Right Math Flash Cards - Division Flash Cards - 156 Hole
    Punched Math Game Flash Cards - 2 Binder Rings - for Ages 8 and Up -
    3rd, 4th, 5th and 6th Grade
[01] Bicycle Dice, 10 Count (Six Sided, 16 mm)
[02] Cards Against Humanity: Absurd Box - 300-Card Expansion
[03] Ultra Pro Silver Series 9 Pocket Pages (25 count pack)
[...]
[19] 5 Surprise Mini Brands Series 2 by ZURU (2 Pack) Amazon Exclusive
    Mystery Real Miniature Brands Collectible Toy Capsule - Gold ( PVC
    Tube Packaging)
```

Real: each shown candidate's own most-voted review

```
[Retrieved information for some candidates]
[00] Star Right Math Flash Cards - Division Flash Cards - 156 Hole
    Punched Math Game Flash Cards - 2 Binder Rings - for Ages 8 and Up -
    3rd, 4th, 5th and 6th Grade
  - 1* "I was pretty impressed with these flash cards. The quality is
      great, they seem like theyll last awhile, and my kids are [...]"
[02] Cards Against Humanity: Absurd Box - 300-Card Expansion
  - 5* "This is my FAVORITE of all the many expansion packs/decks we have
      bought. I got this for my husband for his birthday [...]"
[12] Carson Dellosa 122-Piece American Sign Language Flash Cards for
    Kids, ASL Flash Cards for Toddlers, Learn Sign Language for
    Beginners, Illustrated Signing Game
  - 3* "Card are of good quality. Pictures are good to see detail.<br
      />The box clearly states it is American Sign Language. (ASL) The
      [...]"
[end]
```

Swapped: each heading carries the next candidate's review

```
[Retrieved information for some candidates]
[00] Star Right Math Flash Cards - Division Flash Cards - 156 Hole
    Punched Math Game Flash Cards - 2 Binder Rings - for Ages 8 and Up -
    3rd, 4th, 5th and 6th Grade
  - 5* "This is my FAVORITE of all the many expansion packs/decks we have
      bought. I got this for my husband for his birthday [...]"
[02] Cards Against Humanity: Absurd Box - 300-Card Expansion
  - 3* "Card are of good quality. Pictures are good to see detail.<br
      />The box clearly states it is American Sign Language. (ASL) The
      [...]"
[12] Carson Dellosa 122-Piece American Sign Language Flash Cards for
    Kids, ASL Flash Cards for Toddlers, Learn Sign Language for
    Beginners, Illustrated Signing Game
  - 1* "I was pretty impressed with these flash cards. The quality is
      great, they seem like theyll last awhile, and my kids are [...]"
[end]
```

Template only

```
[Retrieved information for some candidates]
[00] Star Right Math Flash Cards - Division Flash Cards - 156 Hole
    Punched Math Game Flash Cards - 2 Binder Rings - for Ages 8 and Up -
    3rd, 4th, 5th and 6th Grade
  - (no information available)
[02] Cards Against Humanity: Absurd Box - 300-Card Expansion
  - (no information available)
[12] Carson Dellosa 122-Piece American Sign Language Flash Cards for
    Kids, ASL Flash Cards for Toddlers, Learn Sign Language for
    Beginners, Illustrated Signing Game
  - (no information available)
[end]
```

Matched filler: boilerplate of the same word count

```
[Retrieved information for some candidates]
[00] Star Right Math Flash Cards - Division Flash Cards - 156 Hole
```

```
    Punched Math Game Flash Cards - 2 Binder Rings - for Ages 8 and Up -
    3rd, 4th, 5th and 6th Grade
  - this entry was last checked by the catalogue team and the record
      below was filed for internal use only no further details were
      attached to this listing at the time the note is retained here so
      that the layout of each entry stays the same across items and it
      carries nothing about the product itself this entry was last
      checked by the catalogue team and the record below was filed for
[02] Cards Against Humanity: Absurd Box - 300-Card Expansion
  - this entry was last checked by the catalogue team and the record
      below was filed for internal use only no further details were
      attached to this listing at the time the note is retained here so
      that the layout of each entry stays
[12] Carson Dellosa 122-Piece American Sign Language Flash Cards for
    Kids, ASL Flash Cards for Toddlers, Learn Sign Language for
    Beginners, Illustrated Signing Game
  - this entry was last checked by the catalogue team and the record
      below was filed for internal use only no further details were
      attached to this listing at the time the note is retained here so
      that the layout of each entry stays the same across items and it
      carries nothing about the product itself this entry was last
      checked by the catalogue team and the record below was filed for
      internal use only no further details
[end]
```

The four controls (§4). The rule here is the contrastive objective with one shared budget of ten over all twenty candidates; its *allocation* is how many of the ten passages each candidate received. The *count-matched* control draws ten passages uniformly from the union of all twenty pools, so its allocation is whatever the draw gives and its passages are chosen without reference to the request. The *presence-matched* control gives every candidate one passage drawn uniformly from its own pool. The *content-resampled* control keeps the rule's allocation exactly and refills each slot uniformly from that candidate's own pool, so the prompt has the same shape and the same counts and only the text differs. The *swapped* control keeps the rule's own passages and re-files them among the candidates the rule covered, as a derangement, so that no block sits under its owner; the prompt contains exactly the same text. Every draw uses a seed fixed per instance.

The contrastive rule, one shared budget of ten: its allocation

```
allocation: [03]x5, [06]x2, [12]x1, [13]x2
(the target is [03])
```

The rule's own block

```
[Retrieved reviews for the candidates]
[03] Ultra Pro Silver Series 9 Pocket Pages (25 count pack)
  - 5* "I got these sleeves to go in my son's new baseball album, and
      they are so much nicer than any other I [...]"
  - 5* "I ordered extra pages for my grandson's Pokemon trading card
      binder. Unfortunately for me, the binder does not accept extra
      pages; nevertheless, [...]"
  - 4* "I use this for my Pokemon cards its great, the pages have just
      the right thickness to protect each card. I like [...]"
  - 5* "My 7 year old son has developed an obsession with Pokemon cards.
      While he has no clue what they are and what [...]"
  - 5* "This is exactly what I needed! I use this for my Animal Crossing
      amiibo cards and I was tired of just putting [...]"
[06] Joyoldelf Cool Black Foil Poker Playing Cards, Waterproof Deck of
    Cards with Gift Box, Use for Party and Game
  - 5* "These cards were really cool and I liked the fact that they don't
      bend easily unlike traditional paper cards.<br /><br />One
      disadvantage [...]"
  - 5* "These are fun, plastic cards. They hold up well."
[12] Carson Dellosa 122-Piece American Sign Language Flash Cards for
    Kids, ASL Flash Cards for Toddlers, Learn Sign Language for
    Beginners, Illustrated Signing Game
  - 5* "They're simple and clear. Coated cards.<br />The 105 cards are in
```

```
      a sturdy box.<br />Contains ASL alphabet, numbers 1-10, one card
      for [...]"
[13] Ultra PRO 3" x 4" Clear Regular Top Loaders For Cards With Card
    Sleeves Bundle Standard Size 100ct Baseball Card Sleeves Trading Card
    Sleeve Baseball Card Protectors Toploader Card Protectors
  - 5* "Very nice quality sleeves & top loaders!!<br />If you need to put
      so cards in top loaders this is all you need, [...]"
  - 5* "worked perfect with pokemon cards"
[end]
```

Count-matched: ten passages drawn uniformly from all pools

```
allocation: [02]x1, [03]x1, [04]x1, [05]x1, [07]x2, [08]x1, [13]x1,
    [16]x1, [17]x1
[Retrieved reviews for the candidates]
[02] Cards Against Humanity: Absurd Box - 300-Card Expansion
  - 5* "There are some great cards in here, including an epic deja vu
      moment..."
[03] Ultra Pro Silver Series 9 Pocket Pages (25 count pack)
  - 5* "I am actually using this product as coupon organizer inside a
      notebook . They work great and I am totally happy with [...]"
[04] (10 Pack) EcoTEK Protectors Pop Protector Compatible with - 4" Inch
    Funko POP! Figures, Strong, Clear Pop Case, Acid Free
  - 1* "Wow, these are beyond terrible. When you peel off the protective
      film, theres tons of marks and scuffs. So definitely not a [...]"
[05] Crayola Epic Book of Awesome, All-in-One Coloring Book Set, 288
    Pages, Kids Indoor Activities, Gift
  - 1* "I searched for giant meaning large pages not a huge coloring
      book. I guess that is my fault but when it said [...]"
[07] SIQUK 4 Pieces Dice Tray Set Foldable Dice Trays Hexagon Dice
    Rolling Tray PU Leather Dice Holder for Dice Games Like RPG, DND and
    Other Table Games(Orange, Yellow, Cyan, Violet)
  - 5* "These are a great purchase for anyone who rolls dice for any
      game.<br />Pros:<br />- inexpensive<br />- quiet<br />- keeps dice
      contained<br [...]"
  - 5* "Dimensions are given but I wanted to emphasize these are very
      decent sized! You get 4 of these, yet you can pay [...]"
[08] Melissa & Doug Sketch Pad (9 x 12 inches) - 50 Sheets, 2-Pack - Kids
    Drawing Paper, Drawing And Coloring Pad Art Supplies
  - 5* "Love them! Bought the 2 pack for my granddaughter that is almost
      2...one for my daughter to take home and one for [...]"
[13] Ultra PRO 3" x 4" Clear Regular Top Loaders For Cards With Card
    Sleeves Bundle Standard Size 100ct Baseball Card Sleeves Trading Card
    Sleeve Baseball Card Protectors Toploader Card Protectors
  - 5* "These are well made for protecting cards and I would purchase
      again. Arrived on time in perfect condition."
[16] Crayola Baby Shark Color Wonder Pages (18 Pages & 5 Markers), Mess
    Free Coloring For Toddlers, Gift for Kids, Travel Activities
  - 5* "I have a busy 2-year-old and a 8-month-old. So, I'm constantly
      trying to find activities to keep her busy, but I'm usually [...]"
[17] Learning Resources Pretend Play Money - 150 Pieces, Ages 3+ Play
    Money for Kids, Pretend Money for Kids, Play Money Set, Money and
    Banking Play Toys, Toddler Learning Toys
  - 4* "Bought Learning Resources ash register 5 years ago and kids still
      play with it. Held up to a lot of hard play [...]"
[end]
```

Presence-matched: one random passage per candidate (first three shown)

```
[Retrieved reviews for the candidates]
[00] Star Right Math Flash Cards - Division Flash Cards - 156 Hole
    Punched Math Game Flash Cards - 2 Binder Rings - for Ages 8 and Up -
    3rd, 4th, 5th and 6th Grade
  - 1* "You think it would be a simple thing for quality control to
      ensure all the cards were included. Missing 10x9. When ;you [...]"
[01] Bicycle Dice, 10 Count (Six Sided, 16 mm)
  - 1* "Immediately upon using the dice, (and I do know die is plural for
```

```
      dice, but all of you know what I mean), [...]"
[02] Cards Against Humanity: Absurd Box - 300-Card Expansion
  - 5* "Always a good thing to have more card to play with to add in
      extra laughs."
[end]
[... and one under each of the other seventeen]
```

Content-resampled: the rule's allocation, slots refilled from each pool

```
allocation: [03]x5, [06]x2, [12]x1, [13]x2
[Retrieved reviews for the candidates]
[03] Ultra Pro Silver Series 9 Pocket Pages (25 count pack)
  - 3* "Original Rating: 5 Stars. Perfect. The pages I got were from the
      correct brand.<br /><br />Update: I ordered another set of these
      [...]"
  - 5* "I got these sleeves to go in my son's new baseball album, and
      they are so much nicer than any other I [...]"
  - 5* "Make sure you count out how many pages you'll need so you don't
      have to continuously make multiple purchases. My daughter purchased
      [...]"
  - 1* "This was suggested by Amazon for a product that it didn't fit
      in."
  - 1* "The pockets are so tight I can't fit anything in them. Hate it!!"
[06] Joyoldelf Cool Black Foil Poker Playing Cards, Waterproof Deck of
    Cards with Gift Box, Use for Party and Game
  - 5* "I like collecting decks of cards. Not to play with them, I get
      them because I like the design. Black is beautiful [...]"
  - 5* "Love these - take them out by the pool or on the boat to play and
      don't worry! A little slippery, but [...]"
[12] Carson Dellosa 122-Piece American Sign Language Flash Cards for
    Kids, ASL Flash Cards for Toddlers, Learn Sign Language for
    Beginners, Illustrated Signing Game
  - 5* "Perfect!!! Read a lot of the reviews prior to purchasing. Glad I
      made this purchase over others.<br /><br /> bought a hole [...]"
[13] Ultra PRO 3" x 4" Clear Regular Top Loaders For Cards With Card
    Sleeves Bundle Standard Size 100ct Baseball Card Sleeves Trading Card
    Sleeve Baseball Card Protectors Toploader Card Protectors
  - 1* "All the packs I had received had chipped corners or edges on most
      of the card holders. Definitely do not recommend these [...]"
  - 5* "They protect my cards and they were a decent price! I definitely
      recommend "
[end]
```

Swapped: the rule's passages re-filed among the covered candidates

```
re-filing: [03] shows [06]'s, [06] shows [03]'s, [12] shows [13]'s, [13]
    shows [12]'s
[Retrieved reviews for the candidates]
[03] Ultra Pro Silver Series 9 Pocket Pages (25 count pack)
  - 1* "Cant even use these to play with. You literally cannot see what
      card you have even if its right if front of [...]"
  - 5* "my wife loves Poker and the moment I saw some black cards I knew
      she would love it so got it for [...]"
  - 1* "By the description I thought these cards would be black with a
      gold imprint but they are black on black and are [...]"
  - 5* "Used these are a photo prop for a Kakegurui cosplay! They matched
      the aesthetic perfectly and showed up great in photos. Played
      [...]"
  - 5* "My husband and I have friends over occasionally to play cards at
      our house. So upon looking for some new cards I [...]"
[06] Joyoldelf Cool Black Foil Poker Playing Cards, Waterproof Deck of
    Cards with Gift Box, Use for Party and Game
  - 5* "Good for yugioh cards."
  - 2* "Read carefully when you order these, which I unfortunately
      didn't. These are from the Ultra Pro Silver series which is of a
      [...]"
[12] Carson Dellosa 122-Piece American Sign Language Flash Cards for
```

```
    Kids, ASL Flash Cards for Toddlers, Learn Sign Language for
    Beginners, Illustrated Signing Game
  - 5* "Great protector"
[13] Ultra PRO 3" x 4" Clear Regular Top Loaders For Cards With Card
    Sleeves Bundle Standard Size 100ct Baseball Card Sleeves Trading Card
    Sleeve Baseball Card Protectors Toploader Card Protectors
  - 5* "Perfect!!! Read a lot of the reviews prior to purchasing. Glad I
      made this purchase over others.<br /><br /> bought a hole [...]"
  - 3* "Card are of good quality. Pictures are good to see detail.<br
      />The box clearly states it is American Sign Language. (ASL) The
      [...]"
[end]
```

C THE SELECTION ALGORITHMS, AND WHAT INSPIRED THEM

A rule sees the request embedding q, one pool of passage embeddings per candidate, and a budget B. Every embedding is from one small sentence encoder, `all-MiniLM-L6-v2` with 22M parameters (Wang et al., 2020; Sentence-Transformers, 2021; Reimers & Gurevych, 2019), L_2-normalised so that a dot product is a cosine. No rule may see which candidate is the target; the oracle bound is the one exception and is a bound, not a method. Every rule is deterministic given its inputs, and the controls of Table 2 are drawn from a seed fixed per instance.

Four objectives. *Relevance* scores a passage by $\cos(e, q)$, the rule a retrieval-augmented system uses by default. *Contrastive* relevance scores it by $\cos(e, q) - \lambda \max_{j \neq \text{owner}(e)} \cos(e, c_j)$ with c_j the mean embedding of candidate j's pool and λ=0.5: a passage that every candidate's pool could have supplied says nothing about which is the target, so it is charged for how well it matches the competitors. This was inspired by an encoder-side gate we ran before any reader: on the hardest quartile of Amazon instances the contrastive score separates the target from its competitors where plain relevance does not. *MMR* is maximal marginal relevance (Carbonell & Goldstein, 1998), selected greedily, each step taking the passage maximising $\beta \cos(e, q) - \rho \max_{f \in \text{chosen}} \cos(e, f)$ with β=ρ=0.5; this is the original rule with both similarity functions cosine and $\lambda = \beta/(\beta + \rho)$, since the argmax is invariant to a positive rescaling, and at ρ=0 it is relevance. The *determinantal* rule scores a set by the determinant of a kernel with quality $\exp(\beta \cos(e, q)/2)$ on the diagonal and cosine similarities off it, greedily maximised as in Chen et al. (2018), following the determinantal point processes of Kulesza & Taskar (2012): near-duplicates cost it almost everything, so it is the strongest form of the intuition that a set of evidence should be diverse. Because the quality term is the same in MMR and the determinantal rule, the two differ only in whether the redundancy charge is bounded (a maximum) or unbounded (a determinant). Neither λ nor β is tuned: at this sample size the choice of β is not identifiable (15 of 16 rule-by-cell combinations select an unstable value across repeated splits), and the held-out effects at a tuned β lie in the same band as those at 0.5.

Two allocations. Under *one each*, every candidate carries its own best passage under the objective and B equals the number of candidates; the count per candidate is equal by construction, and the presence-matched control is its counterpart. Under a *shared budget*, every passage of every candidate is pooled and the B best are taken, so a candidate may carry several or none; the count-matched control is its counterpart. The one-each determinantal rule is a partition-constrained determinantal process, which couples the choices across candidates even when each contributes one row and so does not collapse to relevance. Crossing the four objectives with the two allocations gives the eight rules of Table 6.

Allocation against content. Pooled over the eight shortlist cells, moving from one passage each to a shared budget is worth +0.021 HR@1 for relevance and +0.025 for the contrastive objective, positive in eight and seven cells of eight; changing the objective under a shared budget is worth +0.005. Four measurements say why. *Counts predict the reader.* Fitting a softmax over the shortlist to the reader's own choices, with the number of passages each candidate received as the only feature, improves the fit by a likelihood ratio of $\chi^2_1 = 82.4$ to 109.2 across the four cells and both shared-budget rules, and by exactly zero for the one-each rules, as it must with flat counts; adding the relevance mass placed before each candidate barely improves on counts alone. *Resampling the text*

under a fixed allocation costs nothing (Table 4). The content-resampled control splits a shared-budget rule's gain over the count-matched control into two parts that add exactly, per instance: the part that came from where the budget went (the resampled control minus the count-matched one) and the part that came from which text filled the slots (the rule minus its resampled control). Both shared-budget rules were run with their controls in one leg per cell, on the twenty-review pools, at $K{=}3$ and 10. Over the sixteen rule-by-cell combinations the allocation part is $+0.008$ to $+0.054$ HR@1, its interval above zero in thirteen and surviving correction in eleven; the content part is -0.012 to $+0.021$, with no interval clear of zero in either direction. Pooled, allocation is worth $+0.032$ and content $+0.005$. The one cell in which the two are of a size is Reddit with metadata at $K{=}10$ for relevance, $+0.024$ against $+0.021$, both intervals touching zero.

Table 4: **Where a shared-budget rule's gain comes from.** Rule minus the count-matched control (total), split into the allocation part (content-resampled control minus count-matched) and the content part (rule minus content-resampled), which add exactly per instance. Twenty-review pools, Qwen3-8B on the reader's own top-K, one leg per cell; **bold** survives Holm within the four parts of a cell, *italic* has an interval that excludes zero without surviving.

K	rule, term	Amazon-20 titles only	Amazon-20 + metadata	Reddit titles only	Reddit + metadata
3	relevance, total	*+0.028*	*+0.028*	+0.021	−0.001
	allocation	**+0.035**	**+0.025**	+0.020	+0.008
	content	−0.007	+0.003	+0.001	−0.009
	contrastive, total	+0.013	*+0.027*	*+0.045*	+0.017
	allocation	*+0.025*	**+0.025**	**+0.029**	**+0.023**
	content	−0.012	+0.001	+0.016	−0.005
10	relevance, total	*+0.040*	*+0.064*	*+0.047*	*+0.045*
	allocation	**+0.040**	**+0.050**	*+0.032*	+0.024
	content	+0.000	+0.014	+0.015	+0.021
	contrastive, total	*+0.041*	*+0.057*	*+0.064*	*+0.053*
	allocation	**+0.035**	**+0.054**	**+0.048**	**+0.036**
	content	+0.007	+0.003	+0.016	+0.017

Designed allocations. *Designing the allocation from the reader's own prior reproduces the gain.* Scoring is held at plain relevance inside every pool and only the split of the K passages varies, set from the reader's own first-stage scores over the shortlist: all to its leader; in proportion to the softmax of its scores; or in proportion to one minus that softmax, which starves the candidate the reader prefers. On Amazon-20 with titles at $K{=}10$ the encoder-driven shared-budget rule is worth $+0.039$ $[+0.012, +0.067]$, the leader allocation $+0.041$ $[+0.014, +0.068]$ and the softmax allocation $+0.048$ $[+0.022, +0.075]$, all surviving correction; the inverse allocation is -0.038 $[-0.065, -0.010]$. Across the eight cells the inverse allocation is harmful in all eight and survives correction in six, by -0.028 to -0.077 (Table 5). A reader that finds nothing under the candidate it already prefers, while the others carry something, demotes it; the shared-budget rules were doing this in reverse. These three rules were designed to test the allocation reading directly, and they are the clearest statement of availability bias working *for* a rule.

Table 5: **Allocations designed from the reader's own prior**, against the presence-matched control, one five-arm family per cell; **bold** survives Holm, *italic* has an interval that excludes zero without surviving.

		K	one each	shared	leader	uncertain	inverse
Amazon-20	titles only	3	*+0.023*	*+0.030*	+0.005	+0.010	**−0.042**
		10	+0.020	**+0.039**	**+0.041**	**+0.048**	**−0.038**
Amazon-20	+ metadata	3	+0.007	*+0.029*	+0.012	+0.015	*−0.029*
		10	**+0.031**	**+0.061**	**+0.048**	**+0.060**	−0.028
Reddit	titles only	3	+0.008	+0.021	+0.023	+0.024	**−0.063**
		10	+0.016	**+0.047**	**+0.047**	**+0.047**	**−0.063**
Reddit	+ metadata	3	−0.004	−0.001	+0.024	+0.023	**−0.052**
		10	+0.001	**+0.045**	*+0.032*	**+0.037**	**−0.077**

The reader as selector. An 8B reasoning model asked to choose the passages itself, the one selector with no encoder in it, is no better than cosine with one passage each, -0.016 $[-0.048, +0.016]$ at $K{=}3$ and -0.028 $[-0.072, +0.012]$ at $K{=}10$; giving that same selector a shared budget is worth $+0.040$ $[+0.008, +0.076]$ and $+0.072$ $[+0.024, +0.120]$ on top, the largest allocation effect we measured ($n{=}250$, Reddit). It agrees with cosine on 0.16 of its picks, so it is not cosine by another name.

Outside the family. Three further selectors were measured at an encoder-side gate, the shortlist that contains the target, one passage per candidate, ranked by the selector's own score, before any reader was run, under the rule that a scorer which loses to cosine before the reader sees it cannot be what the reader is missing. *A cross-encoder* trained for web passage relevance (`ms-marco-MiniLM-L6-v2`, Bajaj et al., 2018) scores no better than cosine on Amazon-20 ($+0.001$ $[-0.035, +0.040]$ at $K{=}3$, $+0.009$ at $K{=}10$) and worse on Reddit (-0.025 and -0.056 $[-0.093, -0.019]$). *An in-domain encoder* contrastively pretrained on Amazon review–metadata pairs (Hou et al., 2024a) moves the gate by less than the contrastive term does in every cell. *A set-level score*, the log-determinant of a candidate's own best subset of passages, ranks candidates better than max-cosine on the deep Reddit pool ($+0.027$ and $+0.040$) but, used to choose which passages to show, is below cosine in seven of eight cells at three passages and survives correction in none: a property that identifies a good candidate is not thereby a property that selects good evidence.

Scope. The four objectives are relevance-shaped functions of one encoder's cosine, and the rules are compared on the reader's own top-K, whose pools all match the request about equally well. Within that family and on that shortlist, allocation dominates: a rule appears to work when it concentrates evidence under the candidates the reader is already inclined towards, and appears not to when it spreads evidence evenly. Three things keep the statement from being broader. The determinantal rule, the one objective that scores the set rather than the passage, gains $+0.001$ from a shared budget, so the shape of the objective decides whether allocation helps at all. On the full list the passage's content is worth more than the same rule's shortlist effect (§6), so the objective does carry signal; it does not separate rules where the candidates are already alike. And none of the selectors above was trained for the task; no learned selector was run. The claim is therefore about what separates the heuristics people actually use, not about the ceiling on selection: the oracle bound of Table 6 is $+0.09$ to $+0.12$ HR@1 over the control at $K{=}10$, and the best rule takes a third to two thirds of it.

The eight rules on the shortlist. Table 6 gives every rule against its matched control in the eight cells behind §6: every shared-budget rule survives correction in several cells, the two diversity objectives survive with one passage each on Reddit at $K{=}10$, the oracle bound is $+0.09$ to $+0.12$ at $K{=}10$, and the presence term never leaves the floor. Three presentation seeds reproduce the shortlist effects (Appendix F).

Table 6: **Eight selection rules against their matched controls**, HR@1, Qwen3-8B on the reader's own top-K, one leg per cell. **Bold** survives Holm within the eight-rule family, *italic* has an interval that excludes zero without surviving.

rule	Amazon-20 titles only		Amazon-20 + metadata		Reddit titles only		Reddit + metadata	
	K=3	K=10	K=3	K=10	K=3	K=10	K=3	K=10
n	734	691	724	670	750	750	750	750
recall@K	0.452	0.699	0.468	0.724	0.533	0.765	0.501	0.757
one each, minus presence-matched								
relevance	*+0.023*	+0.020	+0.007	*+0.031*	+0.008	+0.016	−0.004	+0.001
contrastive	+0.012	*+0.035*	+0.003	+0.010	*+0.024*	+0.020	−0.005	+0.015
MMR	+0.018	+0.014	+0.012	+0.000	*+0.029*	**+0.044**	+0.001	**+0.039**
determinantal	+0.014	+0.022	+0.008	+0.004	*+0.031*	**+0.060**	+0.005	**+0.049**
shared budget, minus count-matched								
relevance	*+0.030*	**+0.039**	*+0.029*	**+0.061**	+0.021	**+0.047**	−0.001	**+0.045**
contrastive	+0.012	**+0.045**	*+0.029*	**+0.045**	**+0.045**	**+0.064**	+0.017	**+0.053**
MMR	*+0.030*	**+0.051**	+0.014	**+0.040**	*+0.031*	**+0.057**	+0.005	**+0.057**
determinantal	+0.016	+0.023	+0.008	*+0.031*	+0.023	**+0.049**	+0.008	**+0.039**
oracle, minus count-matched	*+0.044*	*+0.120*	*+0.029*	*+0.088*	*+0.052*	*+0.121*	*+0.032*	*+0.104*
presence-matched, minus none	+0.008	−0.009	−0.008	−0.016	+0.011	+0.009	+0.013	+0.005

D THE AVAILABILITY EXPERIMENT IN DETAIL

The measurement. The reader's pointwise readout gives one score per candidate, and their softmax over the twenty is the distribution the reader's answer is read from. For a shown set S let $m = \sum_{i \in S} p_i$ be the probability that the answer is one of the shown candidates. Figure 2 and the tables here report the change in $\log \frac{m}{1-m}$, the log-odds, with the evidence against without it on the same instance and presentation, both read in fresh direct sessions, with m clipped to $[10^{-3}, 1-10^{-3}]$. Intervals are percentile bootstraps over instances; contrasts between contents are paired.

Three shown sets. Table 7 gives every content under three shown sets on both corpora: the three candidates the reader's own reasoning read ranks highest, which is the set of §4; three candidates drawn at random; and the reasoning read's bottom three. Figure 4 plots the review columns. The direct read's own top three are not a usable shown set: they already hold 0.96 of its probability and cannot move, whereas the reasoning read's top three hold 0.72 under the direct read. Under every set and on both corpora, real and swapped reviews move the odds the same way, and the paired real-minus-swapped row has an interval containing zero; matched filler moves them part of the way under the top three and not at all under three random candidates, where the bare template loses the most.

Table 7: **The shift of the log-odds that the answer is one of the shown three**, every content under three shown sets, Qwen3-8B, direct read, n=750 per corpus, with 95% intervals and the paired real-minus-swapped contrast per set.

content	Amazon-20, reviews	Amazon-20, metadata	Reddit, reviews	Reddit, metadata
the reader's top three (probability held before evidence 0.72 / 0.73)				
real	+1.839	+3.121	+2.110	+2.374
	[+1.592, +2.096]	[+2.869, +3.383]	[+1.898, +2.331]	[+2.147, +2.600]
swapped	+1.923	+3.080	+2.079	+1.560
	[+1.678, +2.168]	[+2.820, +3.344]	[+1.863, +2.297]	[+1.339, +1.780]
matched filler	+0.541	+0.976	+0.752	+0.515
	[+0.357, +0.723]	[+0.790, +1.168]	[+0.573, +0.938]	[+0.334, +0.702]
template only	−0.408	−1.200	−1.030	−2.315
	[−0.614, −0.204]	[−1.419, −0.986]	[−1.225, −0.826]	[−2.535, −2.083]
real − swapped, paired	−0.084	+0.042	+0.031	+0.814
	[−0.260, +0.088]	[−0.074, +0.155]	[−0.096, +0.163]	[+0.676, +0.950]
three at random (probability held before evidence 0.14 / 0.14)				
real	+0.663	+1.840	+0.814	+0.981
	[+0.464, +0.870]	[+1.599, +2.086]	[+0.634, +1.008]	[+0.806, +1.167]
swapped	+0.595	+2.146	+0.704	+0.149
	[+0.389, +0.799]	[+1.913, +2.386]	[+0.536, +0.885]	[−0.001, +0.311]
matched filler	−0.451	−0.166	−0.178	−0.344
	[−0.612, −0.294]	[−0.319, −0.016]	[−0.307, −0.042]	[−0.474, −0.212]
template only	−0.865	−1.158	−0.981	−1.265
	[−1.022, −0.714]	[−1.325, −0.997]	[−1.120, −0.843]	[−1.422, −1.110]
real − swapped, paired	+0.069	−0.306	+0.110	+0.832
	[−0.112, +0.246]	[−0.490, −0.121]	[−0.012, +0.230]	[+0.683, +0.985]
the reader's bottom three (probability held before evidence 0.01 / 0.01)				
real	+0.406	+1.562	+0.176	+0.061
	[+0.266, +0.550]	[+1.340, +1.798]	[+0.042, +0.321]	[−0.033, +0.159]
swapped	+0.505	+1.577	+0.010	−0.147
	[+0.351, +0.669]	[+1.376, +1.795]	[−0.102, +0.132]	[−0.224, −0.066]
matched filler	−0.270	−0.155	−0.403	−0.425
	[−0.336, −0.204]	[−0.230, −0.079]	[−0.482, −0.324]	[−0.505, −0.347]
template only	−0.382	−0.453	−0.534	−0.565
	[−0.462, −0.304]	[−0.537, −0.370]	[−0.621, −0.451]	[−0.657, −0.479]
real − swapped, paired	−0.099	−0.015	+0.166	+0.207
	[−0.250, +0.049]	[−0.165, +0.142]	[+0.061, +0.278]	[+0.136, +0.286]

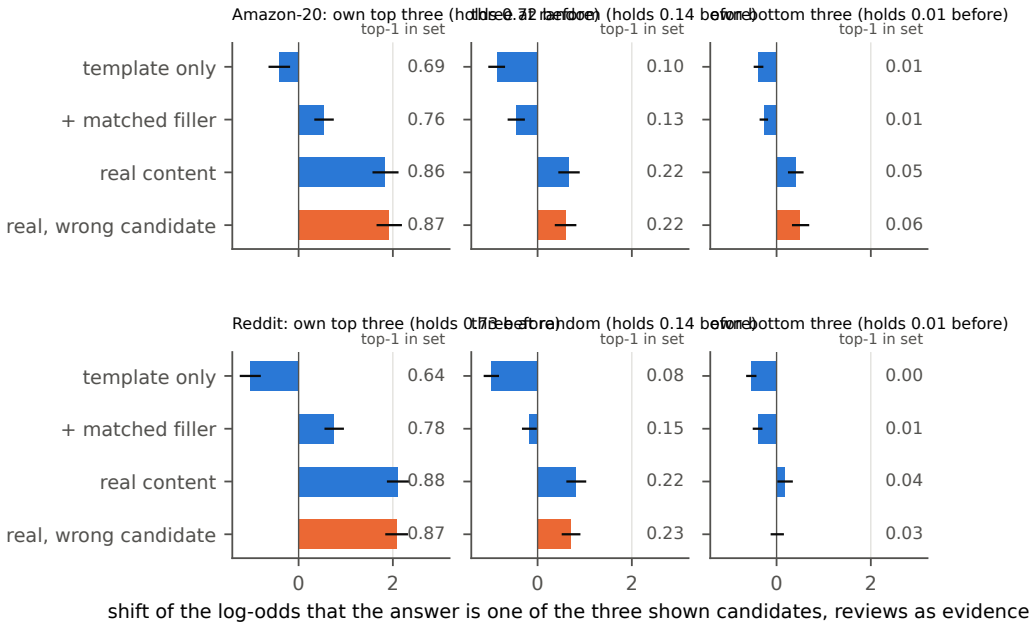

Figure 4: **The same four contents under three shown sets**, reviews as evidence, both corpora; the number in each title is the probability the set holds before evidence, and the right-hand column is how often the top-ranked candidate lands in the set afterwards.

A second reader. Llama-3.1-8B-Instruct has no thinking mode and is read the same way, on the same instances, with the shown sets taken from its own direct read (Table 8); that read is less peaked than Qwen3-8B's, its top three holding 0.56 of its probability, so all three sets have room. It shows every sign of the first reader with larger sizes: real and swapped reviews move the odds the same way under every set on both corpora, filler moves them a fraction of the way, film metadata is verified on Reddit, and under its own bottom three another item's metadata moves enough probability to cut its HR@1 over the twenty from 0.264 to 0.169 on Amazon-20.

Table 8: **Llama-3.1-8B-Instruct as the reader**, the shift of the log-odds that the answer is one of the shown three, every content under its own top three, three at random and its own bottom three, both corpora, direct read, n=750, with 95% intervals.

content	Amazon-20, reviews	Amazon-20, metadata	Reddit, reviews	Reddit, metadata
the reader's top three (probability held before evidence 0.56 / 0.59)				
real	$+0.930$	$+1.835$	$+1.307$	$+1.274$
	$[+0.869, +0.993]$	$[+1.762, +1.910]$	$[+1.248, +1.367]$	$[+1.218, +1.333]$
swapped	$+0.938$	$+1.619$	$+1.215$	$+0.669$
	$[+0.877, +1.001]$	$[+1.546, +1.694]$	$[+1.155, +1.273]$	$[+0.617, +0.721]$
matched filler	$+0.166$	$+0.493$	$+0.239$	$+0.228$
	$[+0.120, +0.213]$	$[+0.442, +0.544]$	$[+0.192, +0.288]$	$[+0.180, +0.279]$
template only	-0.346	-0.251	-0.346	-0.259
	$[-0.392, -0.301]$	$[-0.298, -0.205]$	$[-0.390, -0.301]$	$[-0.305, -0.211]$
real − swapped, paired	-0.009	$+0.216$	$+0.093$	$+0.606$
	$[-0.052, +0.035]$	$[+0.157, +0.275]$	$[+0.048, +0.140]$	$[+0.568, +0.644]$
three at random (probability held before evidence 0.15 / 0.15)				
real	$+0.865$	$+2.209$	$+1.075$	$+0.832$
	$[+0.799, +0.933]$	$[+2.131, +2.290]$	$[+1.001, +1.148]$	$[+0.764, +0.898]$
swapped	$+0.895$	$+2.229$	$+1.067$	$+0.399$
	$[+0.832, +0.960]$	$[+2.149, +2.309]$	$[+0.995, +1.139]$	$[+0.342, +0.453]$
matched filler	-0.163	-0.003	-0.200	-0.275
	$[-0.210, -0.114]$	$[-0.055, +0.053]$	$[-0.251, -0.149]$	$[-0.323, -0.228]$
template only	-0.855	-0.724	-0.955	-0.926
	$[-0.904, -0.805]$	$[-0.774, -0.671]$	$[-1.001, -0.909]$	$[-0.977, -0.876]$
real − swapped, paired	-0.030	-0.019	$+0.008$	$+0.433$
	$[-0.082, +0.020]$	$[-0.096, +0.059]$	$[-0.040, +0.055]$	$[+0.389, +0.477]$
the reader's bottom three (probability held before evidence 0.03 / 0.02)				
real	$+1.188$	$+3.003$	$+1.503$	$+1.043$
	$[+1.107, +1.270]$	$[+2.904, +3.103]$	$[+1.404, +1.599]$	$[+0.963, +1.125]$
swapped	$+1.292$	$+3.075$	$+1.526$	$+0.766$
	$[+1.205, +1.380]$	$[+2.980, +3.175]$	$[+1.431, +1.621]$	$[+0.696, +0.837]$
matched filler	-0.114	$+0.068$	-0.284	-0.271
	$[-0.159, -0.066]$	$[+0.017, +0.122]$	$[-0.330, -0.237]$	$[-0.315, -0.227]$
template only	-0.746	-0.605	-0.900	-0.883
	$[-0.786, -0.704]$	$[-0.649, -0.560]$	$[-0.935, -0.862]$	$[-0.921, -0.843]$
real − swapped, paired	-0.104	-0.072	-0.023	$+0.277$
	$[-0.179, -0.030]$	$[-0.172, +0.023]$	$[-0.092, +0.047]$	$[+0.228, +0.328]$

Dose. Two and four reviews per block instead of one, reviews only, Qwen3-8B on Amazon-20, with the shown sets of the direct read so that the doses pair per instance (Table 9, on the probability scale). Under three random candidates the real shift is $+0.074$, $+0.087$ and $+0.120$ $[+0.098, +0.143]$ of probability at one, two and four reviews and the swapped shift $+0.067$, $+0.098$ and $+0.134$ $[+0.109, +0.159]$; attribution is $+0.006$, -0.011 and -0.014 $[-0.034, +0.007]$. Matched filler, whose length grows with the dose, is -0.019, -0.012 and -0.000; the template, which does not grow, is -0.051 at every dose. More evidence-like text moves more probability, and whose text it is never matters.

Table 9: **One, two and four reviews per block**, the shift of probability onto the shown set, reviews as evidence, Qwen3-8B, Amazon-20, n=750, with 95% intervals; the shown sets are those of the direct read (its top three hold 0.96 and cannot move).

content	1 review per block	2 reviews per block	4 reviews per block
three at random			
real	+0.074 [+0.054, +0.095]	+0.087 [+0.066, +0.108]	+0.120 [+0.098, +0.143]
swapped	+0.067 [+0.047, +0.088]	+0.098 [+0.075, +0.121]	+0.134 [+0.109, +0.159]
matched filler	−0.019 [−0.034, −0.005]	−0.012 [−0.027, +0.002]	−0.000 [−0.016, +0.015]
template only	−0.051 [−0.066, −0.038]	−0.051 [−0.066, −0.038]	−0.051 [−0.066, −0.038]
the reader's top three			
real	+0.007 [−0.003, +0.015]	+0.010 [+0.001, +0.019]	+0.020 [+0.013, +0.028]
swapped	+0.005 [−0.005, +0.014]	+0.013 [+0.004, +0.022]	+0.020 [+0.012, +0.028]
matched filler	−0.037 [−0.049, −0.025]	−0.040 [−0.053, −0.028]	−0.044 [−0.057, −0.032]
template only	−0.101 [−0.118, −0.085]	−0.101 [−0.118, −0.085]	−0.101 [−0.118, −0.084]
the reader's bottom three			
real	+0.038 [+0.026, +0.051]	+0.042 [+0.030, +0.055]	+0.062 [+0.047, +0.078]
swapped	+0.038 [+0.026, +0.051]	+0.044 [+0.031, +0.057]	+0.062 [+0.048, +0.078]
matched filler	−0.000 [−0.000, −0.000]	−0.000 [−0.000, −0.000]	−0.000 [−0.000, +0.000]
template only	−0.000 [−0.001, −0.000]	−0.000 [−0.001, −0.000]	−0.000 [−0.001, −0.000]

Fabrication. A review generated by a Llama-3.1-8B model to assert that a chosen non-target satisfies the request, attached to that candidate alone under the reasoning read's top three, shifts +0.480 [+0.446, +0.513] of probability onto it against +0.064 [+0.042, +0.086] for its own real review, and makes it the reader's first choice on 0.749 of instances against 0.286 before any evidence (n=692); attached to the top three in place of their real reviews, fabricated blocks shift +0.281 [+0.253, +0.310] against +0.133 for the real ones (n=660).

E RETRIEVAL BESIDE REASONING, IN FULL

Six metrics. Table 1 gives HR@1, HR@10, MRR and NDCG@10 with prompt tokens, generated tokens and measured seconds for every configuration in all four settings. The paired contrasts behind §3 are in Table 12: retrieval against the same read without it, retrieval against the count-matched control, reasoning with and without evidence, retrieval alone against reasoning alone, and retrieval before the read against retrieval after reasoning. Retrieval alone minus reasoning alone is +0.016 [−0.017, +0.050], +0.020 [−0.012, +0.052], −0.007 [−0.041, +0.028] and +0.009 [−0.028, +0.045] HR@1 in the four settings; on NDCG@10 it is +0.039 [+0.013, +0.065] on Amazon-20 with titles and inside the floor elsewhere. Reasoning on top of retrieval adds +0.032 [+0.008, +0.056], +0.054 [+0.025, +0.082], +0.037 [+0.009, +0.065] and +0.039 [+0.013, +0.065].

Table 10: **Retrieval after reasoning beside retrieval before the read**, six metrics and cost, Qwen3-8B, budget ten, contrastive selection. The pipeline's tokens include its first stage, and its second read is charged on every instance, as a deployment would pay it.

configuration	HR@1	HR@10	MRR	NDCG@10	prompt	generated	s/query
Amazon-20, titles only ($n = 746$)							
retrieval, no reasoning	0.339	0.779	0.476	0.536	2,836	0	0.27
retrieval and reasoning	0.371	0.775	0.494	0.548	2,836	2,717	43.47
retrieval after reasoning	0.330	0.725	0.462	0.510	3,437	2,909	56.51
retrieval after reasoning, then reasoning	0.354	0.725	0.474	0.518	3,437	5,287	102.03
Amazon-20, titles + metadata ($n = 746$)							
retrieval, no reasoning	0.328	0.768	0.465	0.524	7,429	0	–
retrieval and reasoning	0.382	0.816	0.513	0.575	7,429	2,868	–
retrieval after reasoning	0.324	0.753	0.470	0.524	10,350	3,294	–
retrieval after reasoning, then reasoning	0.343	0.753	0.479	0.531	10,350	5,978	–
Reddit, titles only ($n = 750$)							
retrieval, no reasoning	0.368	0.789	0.509	0.565	1,384	0	0.22
retrieval and reasoning	0.405	0.803	0.534	0.588	1,384	2,569	45.89
retrieval after reasoning	0.371	0.765	0.501	0.552	1,612	2,709	59.30
retrieval after reasoning, then reasoning	0.385	0.765	0.509	0.558	1,612	4,806	93.72
Reddit, titles + metadata ($n = 750$)							
retrieval, no reasoning	0.367	0.788	0.498	0.555	3,284	0	–
retrieval and reasoning	0.405	0.787	0.528	0.578	3,284	2,723	–
retrieval after reasoning	0.353	0.757	0.489	0.541	4,480	3,035	–
retrieval after reasoning, then reasoning	0.385	0.757	0.505	0.552	4,480	5,433	–

Arbitrary evidence. Ten arbitrary reviews add 732 to 1,370 prompt tokens per query over no evidence and forty add 1,798 to 2,916, in the four settings; forty arbitrary reviews reach 0.284, 0.268, 0.316 and 0.304 HR@1 against 0.339, 0.328, 0.368 and 0.367 for ten selected ones. Paired over instances, arbitrary evidence against none is -0.040 $[-0.067, -0.015]$ and -0.046 $[-0.074, -0.019]$ at three and five on Amazon-20 with titles and -0.028 $[-0.055, -0.001]$ in both Reddit-with-metadata cells, each surviving correction; -0.028 $[-0.056, +0.001]$, -0.011, -0.011 and -0.015 at ten; within ± 0.02 at twenty; and $+0.052$ $[+0.023, +0.081]$ and $+0.028$ $[+0.001, +0.055]$ on Reddit at forty against $+0.001$ and $+0.004$ on Amazon. Under a reasoning read ten arbitrary passages are worth -0.047 $[-0.076, -0.017]$, $+0.016$, -0.025 and -0.024, and twenty are worth -0.032 $[-0.064, -0.000]$ on Amazon-20 and -0.044 $[-0.073, -0.015]$ on Reddit with titles only.

Retrieval after reasoning. Table 10 puts the pipeline beside its counterparts. Against retrieval before the read at the same budget and readout, paired on the same instances, it is behind or level in every one of eight comparisons: $+0.009$, $+0.004$, -0.003 and $+0.013$ HR@1 with a direct re-read and $+0.017$, $+0.039$ $[+0.012, +0.066]$, $+0.020$ and $+0.020$ with a reasoning re-read, every interval but one containing zero and none favouring the pipeline; reasoning twice generates 4,806 to 5,978 tokens against 2,569 to 2,868, and the shortlist caps HR@10 at the first stage's recall. On its top ten, ten arbitrary passages are worth -0.008 $[-0.033, +0.017]$, -0.001, -0.004 and $+0.016$ over no evidence in the four settings at 551 to 963 more prompt tokens; the same ten retrieved passages re-filed among the covered candidates are worth -0.032 $[-0.060, -0.004]$, -0.001, -0.071 $[-0.100, -0.041]$ and -0.031, and correct attribution $+0.069$ $[+0.040, +0.097]$, $+0.043$, $+0.144$ $[+0.111, +0.177]$ and $+0.089$, surviving correction in every setting. On a shortlist of three with a budget of three the swap costs $+0.005$ to $+0.057$ and the re-filed passages are within ± 0.02 of none. Before the read, on all twenty, the re-filed passages are worth -0.020, $+0.008$, -0.061 $[-0.093, -0.029]$ and -0.051 and attribution $+0.076$ $[+0.046, +0.109]$, $+0.055$, $+0.165$ $[+0.131, +0.200]$ and $+0.141$ (Table 11 for the titles settings). On the shortlist, where every candidate is plausible, the re-filed passages are worse than none on both corpora, and the cost of misattribution grows with the shortlist and the budget: the pipeline inherits the bias of §4 and depends more than the single read on the retriever's attribution being right on every passage.

Table 11: **The four arms on the reasoning first stage's top ten, and before a read of all twenty**, Qwen3-8B, direct read, budget ten, titles only, one leg per cell. On the shortlist HR@1 counts a target off it as a miss for every arm. Contrasts are paired with 95% intervals; **bold** survives Holm within the five contrasts of a cell, *italic* has an interval that excludes zero without surviving.

	Amazon-20		Reddit	
	after reasoning, K=10	before the read	after reasoning, K=10	before the read
HR@1				
no evidence	0.292	0.283	0.297	0.264
arbitrary	0.284	0.255	0.293	0.253
retrieved	0.329	0.339	0.371	0.368
re-filed	0.260	0.263	0.227	0.203
prompt tokens of the read				
no evidence	1,004	1,370	426	554
arbitrary	1,967	2,733	977	1,288
retrieved / re-filed	2,066	2,836	1,058	1,384
paired contrasts				
retrieved − arbitrary	**+0.045** [+0.016, +0.075]	**+0.084** [+0.052, +0.117]	**+0.077** [+0.047, +0.109]	**+0.115** [+0.083, +0.148]
arbitrary − none	−0.008 [−0.033, +0.017]	−0.028 [−0.056, +0.001]	−0.004 [−0.031, +0.023]	−0.011 [−0.040, +0.019]
retrieved − none	**+0.037** [+0.009, +0.065]	**+0.056** [+0.023, +0.090]	**+0.073** [+0.043, +0.104]	**+0.104** [+0.071, +0.137]
re-filed − none	*−0.032* [−0.060, −0.004]	−0.020 [−0.051, +0.011]	**−0.071** [−0.100, −0.041]	**−0.061** [−0.093, −0.029]
retrieved − re-filed	**+0.069** [+0.040, +0.097]	**+0.076** [+0.046, +0.109]	**+0.144** [+0.111, +0.177]	**+0.165** [+0.131, +0.200]

Table 12: **Paired contrasts at a budget of ten**, HR@1, Qwen3-8B, contrastive selection, with 95% intervals. The row in bold compares the two ways of spending: retrieval with no reasoning against reasoning with no retrieval.

contrast	Amazon-20 titles only	Amazon-20 + metadata	Reddit titles only	Reddit + metadata
retrieval, no reasoning − neither	+0.056 [+0.023, +0.090]	+0.064 [+0.034, +0.095]	+0.104 [+0.071, +0.137]	+0.091 [+0.056, +0.124]
against arbitrary retrieval	+0.084 [+0.052, +0.117]	+0.075 [+0.044, +0.106]	+0.115 [+0.083, +0.148]	+0.105 [+0.072, +0.139]
retrieval and reasoning − reasoning only	+0.048 [+0.013, +0.083]	+0.074 [+0.043, +0.105]	+0.031 [−0.004, +0.065]	+0.048 [+0.015, +0.081]
reasoning, no retrieval − neither	+0.040 [+0.012, +0.068]	+0.044 [+0.015, +0.074]	+0.111 [+0.079, +0.144]	+0.081 [+0.051, +0.112]
retrieval and reasoning − retrieval only	+0.032 [+0.008, +0.056]	+0.054 [+0.025, +0.082]	+0.037 [+0.009, +0.065]	+0.039 [+0.013, +0.065]
retrieval only − reasoning only	+0.016 [−0.017, +0.050]	+0.020 [−0.012, +0.052]	−0.007 [−0.041, +0.028]	+0.009 [−0.028, +0.045]
retrieval only − retrieval after reasoning	+0.009 [−0.020, +0.040]	+0.004 [−0.025, +0.035]	−0.003 [−0.032, +0.027]	+0.013 [−0.019, +0.044]
retrieval and reasoning − retrieval after reasoning, then reasoning	+0.017 [−0.011, +0.046]	+0.039 [+0.012, +0.066]	+0.020 [−0.009, +0.051]	+0.020 [−0.008, +0.048]

Reader size. Table 13 gives the four levels for Qwen3-1.7B, Qwen3-4B and Qwen3-8B. On Amazon-20 with titles, retrieval is worth +0.086 [+0.052, +0.119], +0.068 [+0.034, +0.103] and +0.056 [+0.023, +0.090] HR@1 at the three sizes, and a reasoning trace on an empty prompt +0.017 [−0.009, +0.044], +0.032 [+0.001, +0.062] and +0.040 [+0.012, +0.068]; retrieval alone minus reasoning alone is +0.068 [+0.036, +0.101], +0.036 [+0.003, +0.071] and +0.016 [−0.017, +0.050]. Every point is a pipeline leg, so the presentation order is shared and the four levels are paired on the same instances.

Table 13: **The two levers across readers**, budget of ten, contrastive selection, titles only. HR@10 beside HR@1 because the pipeline's shortlist caps it by construction.

	1.7B		4B		8B	
configuration	HR@1	HR@10	HR@1	HR@10	HR@1	HR@10
Amazon-20, titles only						
no retrieval, no reasoning	0.193	0.646	0.252	0.684	0.283	0.718
retrieval, no reasoning	0.279	0.696	0.320	0.723	0.339	0.779
reasoning, no retrieval	0.210	0.682	0.284	0.682	0.323	0.724
retrieval and reasoning	0.276	0.718	0.363	0.747	0.371	0.775
retrieval after reasoning, then reasoning	0.263	0.689	0.326	0.729	0.354	0.725
Reddit, titles only						
no retrieval, no reasoning	0.113	0.599	0.231	0.668	0.264	0.748
retrieval, no reasoning	0.256	0.684	0.313	0.705	0.368	0.789
reasoning, no retrieval	0.149	0.615	0.307	0.728	0.375	0.764
retrieval and reasoning	0.281	0.688	0.351	0.713	0.405	0.803
retrieval after reasoning, then reasoning	0.251	0.588	0.371	0.723	0.385	0.765

Thinking budget. The reasoning reads above stop the trace at $4{,}096$ generated tokens, which binds on 19 to 21% of reads on Amazon-20 with titles. Re-running both arms with the budget at $8{,}192$ leaves HR@1 where it was: 0.322 without evidence and 0.369 with it against 0.322 and 0.370 under the cap, a paired difference of $+0.000$ $[-0.004, +0.004]$ and -0.001 $[-0.004, +0.000]$ on 743 instances; the mean trace grows from $2{,}879$ to $3{,}262$ tokens without evidence and from $2{,}714$ to $3{,}159$ with it, and 6 to 9% of reads still reach the higher cap. Under the uncapped trace, retrieved evidence is worth $+0.047$ $[+0.012, +0.081]$, and reasoning over the retrieved evidence beats the direct read of the same evidence by $+0.031$ $[+0.005, +0.055]$.

Latency. Twenty queries per corpus, titles only, timed on one idle H100 in bfloat16 at batch size one with the configurations interleaved per query, two warm-up queries discarded and `torch.cuda.synchronize()` around every read; the arbitrary-evidence configurations were timed in a second session of the same design, in which the no-evidence and retrieved reads reproduced their medians within 0.01 seconds. Selection itself takes 0.04 seconds for twenty candidates and 0.02 for ten on CPU. The absolute numbers depend on batching and serving stack; the ratios are the claim.

F Further tables

The full list. Table 14 runs the rules on all twenty candidates, so that coverage is uniform under one-each rules and availability has nothing to favour. Every candidate receives a passage, the reader re-ranks the full list in a fresh direct read, and the no-evidence arm reproduces the first stage's own direct-read HR@1 within the floor (-0.025 to $+0.004$), which is the check that the two prompts are the same prompt. Two budgets are run: one passage per candidate or a shared budget of twenty, and a shared total of ten with ten arbitrary passages as its control. The presence term is -0.023 to $+0.012$; the most-upvoted review, which is the "real" block of §4 chosen with no reference to the request, is within the floor in three cells of four ($+0.044$ on Reddit with titles); ten arbitrary passages are -0.027 to $+0.002$.

Table 14: **Selection on the full list of twenty**, every candidate covered. Left: one passage per candidate or a shared budget of twenty, against the presence-matched control; right: a shared total of ten against ten arbitrary passages. Entries are HR@1 and rule minus control; **bold** survives Holm within the family shown, *italic* has an interval that excludes zero without surviving.

		Amazon-20, titles only		Amazon-20, + metadata		Reddit, titles only		Reddit, + metadata	
budget	arm	HR@1	−rand	HR@1	−rand	HR@1	−rand	HR@1	−rand
one each (20)	none	0.285	+0.023	0.263	−0.003	0.264	−0.008	0.276	−0.012
	rand	0.262	+0.000	0.267	+0.000	0.272	+0.000	0.288	+0.000
	topvote	0.272	+0.010	0.255	−0.012	0.316	**+0.044**	0.285	−0.003
	cosine	0.323	**+0.061**	0.273	+0.007	0.323	**+0.051**	0.303	+0.015
shared, 20	global_topb	0.334	**+0.073**	0.303	*+0.036*	0.348	**+0.076**	0.352	**+0.064**
	global_contrast	0.338	**+0.076**	0.315	**+0.048**	0.356	**+0.084**	0.351	**+0.063**
	dpp	0.281	+0.020	0.276	+0.010	0.352	**+0.080**	0.344	**+0.056**
	mmr_q	0.303	*+0.041*	0.283	+0.017	0.375	**+0.103**	0.357	**+0.069**
shared, 10 total	global_random	0.263	+0.002	0.252	−0.015	0.253	−0.019	0.261	−0.027
	global_topb	0.325	**+0.063**	0.325	**+0.058**	0.339	**+0.067**	0.319	+0.031
	global_contrast	0.341	**+0.079**	0.325	**+0.058**	0.368	**+0.096**	0.367	**+0.079**
	mmr_q	0.300	*+0.038*	0.291	+0.025	0.381	**+0.109**	0.345	**+0.057**

Presentation seeds. The order in which candidates appear moves HR@1, so every comparison holds it fixed across configurations by seeding the shuffle per instance. Re-running the shortlist grid at three presentation seeds, each instance contributing its mean over seeds, MMR's effect over the presence-matched control is stable in all eight cells and the ordering of the shared-budget rules is unchanged (Table 15). The retrieval-before against retrieval-after-reasoning comparison of Appendix E at three seeds is $+0.005$ $[-0.017, +0.027]$, -0.007 $[-0.026, +0.013]$, -0.007 $[-0.028, +0.015]$ and -0.003 $[-0.025, +0.020]$, within the floor in all four settings, with individual seeds on both sides of zero.

Table 15: Rule minus presence-matched control at three presentation seeds, same instances; **bold** survives Holm, *italic* has an interval that excludes zero without surviving.

		K	rule	seed 0	seed 1	seed 2	averaged	spread	stable
Amazon-20	titles only	3	cosine	*+0.023*	*+0.033*	+0.008	*+0.021*	0.025	yes
			dpp	+0.016	*+0.045*	*+0.029*	*+0.030*	0.029	yes
			mmr_q	*+0.030*	*+0.048*	+0.020	*+0.033*	0.027	yes
Amazon-20	titles only	10	cosine	+0.020	*+0.032*	+0.022	*+0.025*	0.012	yes
			dpp	+0.023	+0.019	+0.022	*+0.021*	0.004	yes
			mmr_q	*+0.051*	*+0.045*	*+0.052*	*+0.049*	0.007	yes
Amazon-20	+ metadata	3	cosine	+0.007	+0.017	+0.011	+0.012	0.010	yes
			dpp	+0.008	*+0.039*	+0.010	*+0.019*	0.030	NO
			mmr_q	+0.014	*+0.036*	+0.018	*+0.023*	0.022	yes
Amazon-20	+ metadata	10	cosine	*+0.031*	+0.024	*+0.033*	*+0.029*	0.009	yes
			dpp	*+0.031*	*+0.030*	*+0.057*	*+0.039*	0.027	yes
			mmr_q	*+0.040*	*+0.033*	*+0.043*	*+0.039*	0.010	yes
Reddit	titles only	3	cosine	+0.008	*+0.020*	−0.001	+0.009	0.021	NO
			dpp	+0.023	*+0.023*	+0.008	*+0.018*	0.015	yes
			mmr_q	*+0.031*	*+0.028*	+0.013	*+0.024*	0.017	yes
Reddit	titles only	10	cosine	+0.016	+0.025	*+0.037*	*+0.026*	0.021	yes
			dpp	*+0.049*	*+0.044*	*+0.051*	*+0.048*	0.007	yes
			mmr_q	*+0.057*	*+0.047*	*+0.064*	*+0.056*	0.017	yes
Reddit	+ metadata	3	cosine	−0.004	+0.012	+0.004	+0.004	0.016	NO
			dpp	+0.008	*+0.027*	+0.015	*+0.016*	0.019	yes
			mmr_q	+0.005	*+0.020*	+0.016	*+0.014*	0.015	yes
Reddit	+ metadata	10	cosine	+0.001	+0.019	*+0.032*	+0.017	0.031	NO
			dpp	*+0.039*	*+0.039*	*+0.051*	*+0.043*	0.012	yes
			mmr_q	*+0.057*	*+0.048*	*+0.051*	*+0.052*	0.009	yes

G THE RELEASED IMPLEMENTATION

The controls, the contrast and the family correction are released as a self-contained package (`rerank_min/`) that depends on NumPy alone and runs without a GPU: `controls.py` holds the five controls of Table 2, each with a test asserting that it holds fixed exactly what it claims (the presence-matched control gives every candidate one passage; the count-matched control conserves the total and is free to place several under one candidate; the content-resampled control reproduces the rule's allocation exactly; the swapped control is a derangement of the shown blocks); `protocol.py` holds the paired contrast, the Holm family and the floor; `select.py` holds the eight rules and the designed allocations, with tests for budget conservation, independence from the target and determinism; `merge.py` holds the pipeline's merge rule and the identity $\mathrm{HR@1} = \mathrm{recall} \times \mathrm{decision}$. The research code that produced every number here calls the same functions.