RETRIEVAL EVALUATION

How much policy text is enough?

24 / 24

Two retrieved passages included the intended policy for every question in this benchmark.

Buyer decision

Choose the smallest tested payload that meets the team’s target for retrieving the right policy.

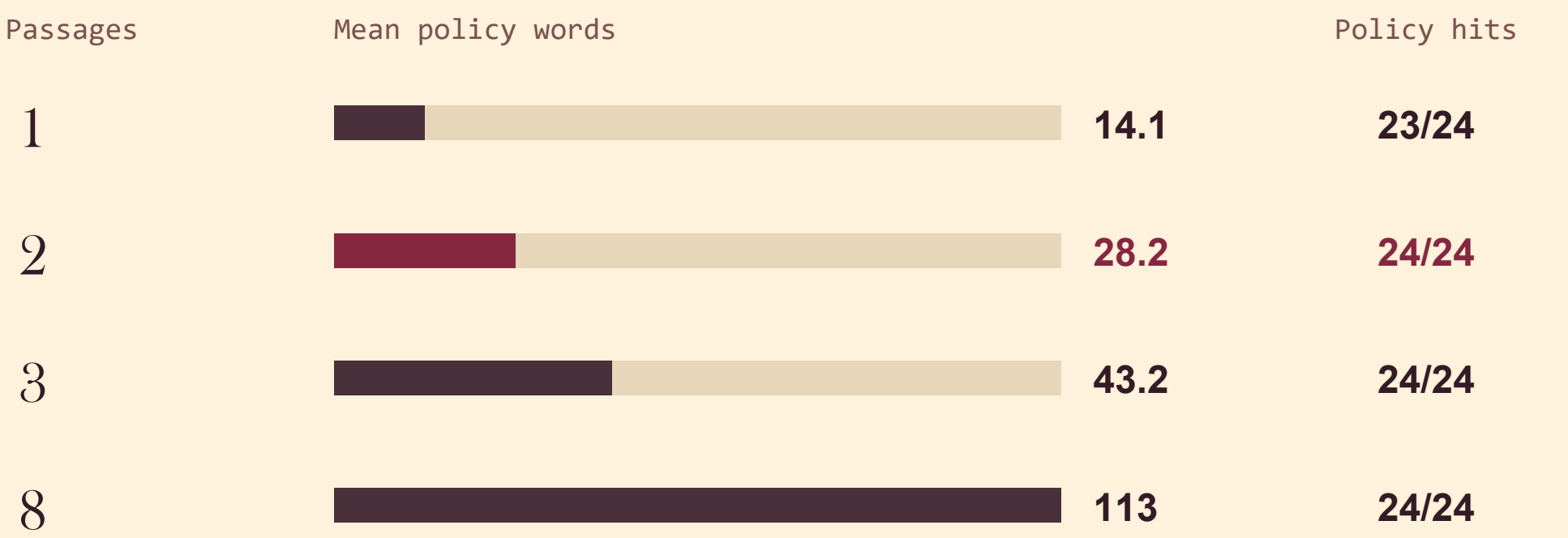
Analysis scope

Eight authored policies, 24 in-scope questions. At two passages the mean payload was 28.2 policy words.

BUDGET SWEEP

Passage count versus policy coverage

Four retrieval budgets are compared on the same 24 questions. Gray tracks share a 113-word maximum.



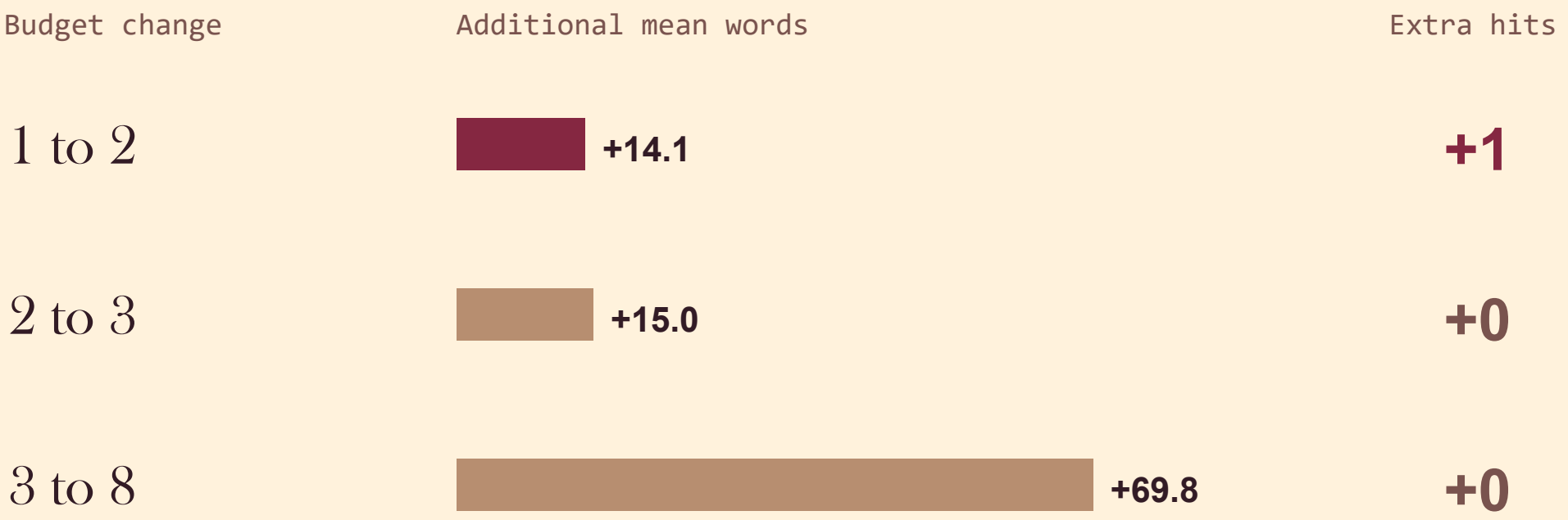
Two passages The smallest tested budget that included the intended policy on every question.

Eight authored policies / 24 in-scope questions / Whitespace-delimited policy words

MARGINAL RETURN

What did the next passage buy?

Each step compares consecutive tested budgets. The final step adds five passages at once.



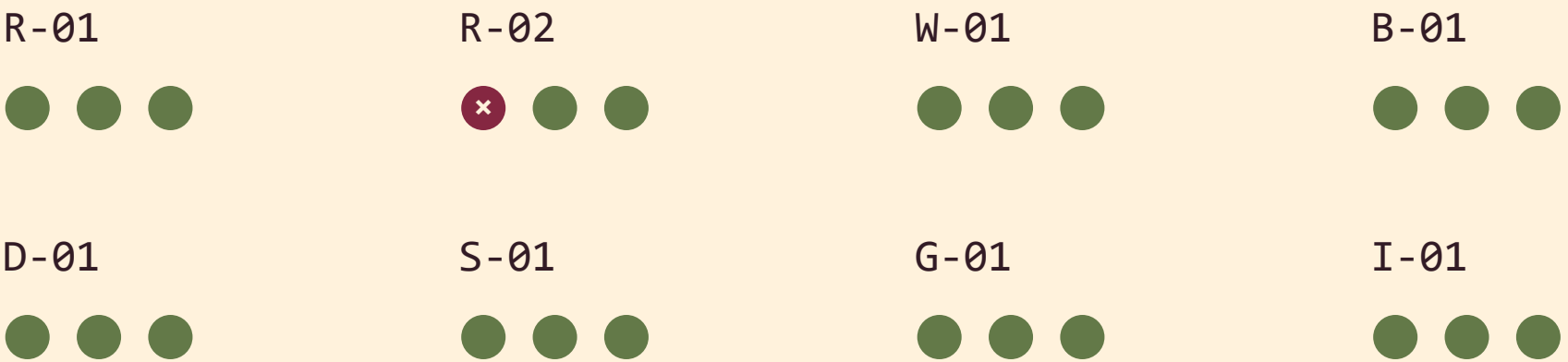
Moving from one passage to two recovered one missed policy. Moving from two to eight added 84.8 words per question with the same observed coverage.

Eight authored policies / 24 in-scope questions / Whitespace-delimited policy words

ERROR ANALYSIS

One question found the wrong policy

The top-one retrieval result is shown for each authored policy group. Each group contains three questions.



The missed question

“I opened it yesterday. Can it go back?”

`Expected R-02  /  Retrieved R-01`

Correct first passage Wrong first passage

Eight authored policies / 24 in-scope questions / Whitespace-delimited policy words

BENCHMARK DESIGN

Choose a budget your team can defend

A retrieval hit means the intended policy appeared in the selected passages. Downstream answer quality is a separate evaluation.

Build a labeled question set

Assign one intended policy to each question. Keep question phrasing distinct from policy text and include hard examples near decision boundaries.

Sweep and count

Run budgets of 1, 2, 3 and 8 on the same 24 questions. Count intended-policy inclusion and average the selected policy words per question.

Pick the operating point

Two passages met full observed coverage at 28.2 words. The one-passage setting used half as many words but missed one policy.

Validate in deployment

Use representative new questions. Measure actual model tokens, latency, answer correctness and billing before projecting production savings.

Source scope / Eight authored policy passages / Instructions and shopper text excluded from word counts