

# Evidence position in English and Spanish

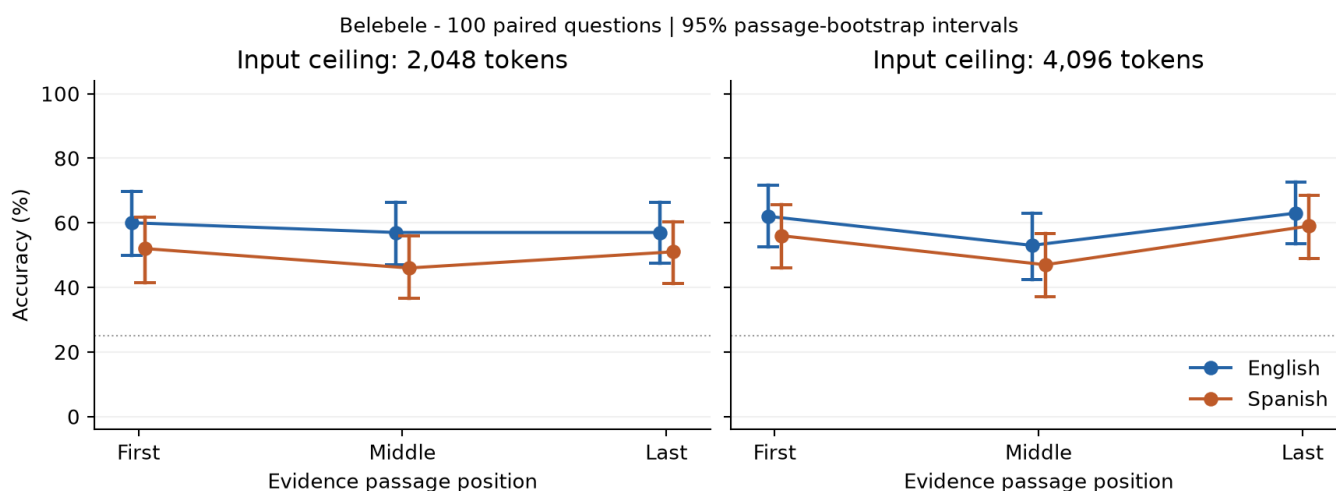
1,600/1,600 cases; 100 paired questions; 96 passage clusters; 0 invalid outputs

## Question and matched design

The complete matched design contains 100 English-Spanish question pairs from 96 distinct passages, and 1,600 committed predictions. Every pair contributes three evidence positions at each of two input ceilings in both languages, plus one no-context and one gold-passage-only response per language. The correct passage is supplied, so this measures evidence use rather than retrieval quality.

A single pinned Qwen3-1.7B MLX 4-bit model is evaluated with thinking disabled through the chat template, greedy decoding, a fresh full context cache for every case, and a one-token A/B/C/D logits mask. Gold-option slots are balanced over the seeded sample; the same option permutation and distractor order are retained across matched conditions.

## Accuracy by evidence position

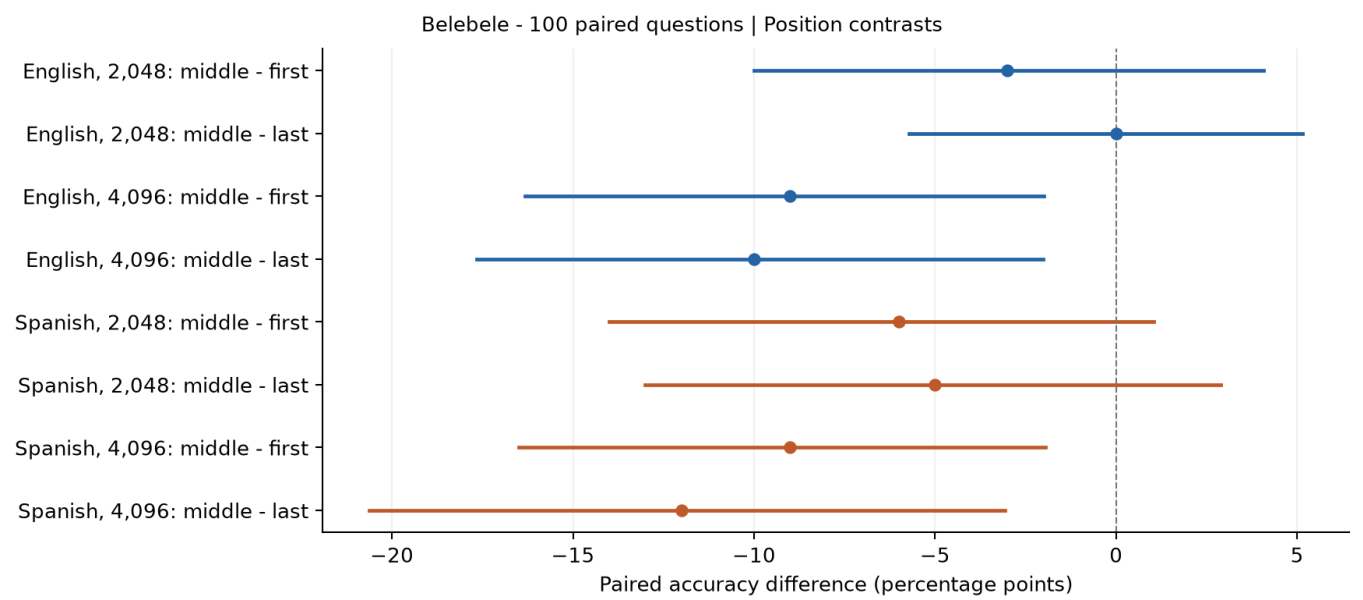


At the 2,048-token ceiling, first/middle/last accuracies were English 60.0%/57.0%/57.0%; Spanish 52.0%/46.0%/51.0%. At the 4,096-token ceiling, first/middle/last accuracies were English 62.0%/53.0%/63.0%; Spanish 56.0%/47.0%/59.0%.

All estimates are question-weighted. The 2,000-draw bootstrap resamples whole passages with replacement using seed 2026; every associated question and matched response remains together. Intervals are 95% percentiles, not adjusted for multiple contrasts. A null interval means fewer than two passage clusters; a collapsed interval does not establish certainty or equivalence.

## Paired position effects and language interaction

The contrast plots show the direction and uncertainty of within-question differences. Negative middle-minus-edge values indicate lower middle-position accuracy in this selected sample. The language-by-position contrast is (Spanish middle - Spanish edge) - (English middle - English edge); it keeps all four responses paired. Its sign describes a relative difference, not an intrinsic property of either language. These are descriptive estimates, not corrected significance tests.

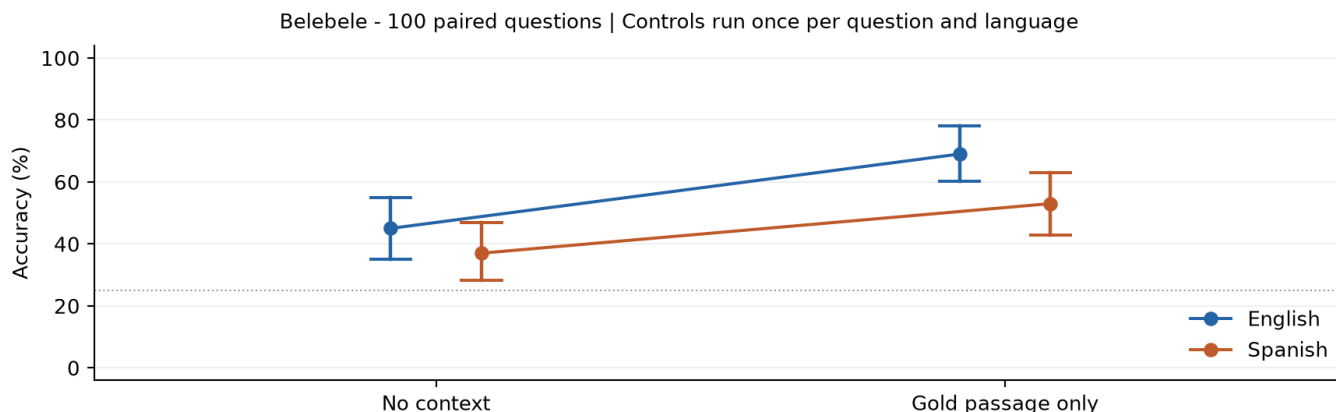


### Language-by-position differences in differences

| Input ceiling | Edge comparison    | Difference, pp [95% interval] |
|---------------|--------------------|-------------------------------|
| 2,048         | Middle minus first | -3.0 [-13.1, +7.7]            |
| 2,048         | Middle minus last  | -5.0 [-14.0, +4.2]            |
| 4,096         | Middle minus first | +0.0 [-9.9, +9.9]             |
| 4,096         | Middle minus last  | -2.0 [-12.7, +8.1]            |

Positive values mean the middle-minus-edge contrast is larger in Spanish. Negative values mean it is smaller. These four contrasts preserve all four within-question responses in each passage resample. Complete Spanish-minus-English and context-minus-gold-only tables accompany the report in CSV and Markdown.

## Controls and actual prompt lengths



The controls are reused in comparisons but their responses are counted once. Gold-only accuracy provides a comprehension reference; no-context accuracy reflects answer selection without supplied evidence. Neither high nor low no-context accuracy establishes the presence or absence of pretraining contamination. The dotted 25% reference is uniform four-choice chance, not an empirical baseline model.

### Recorded input lengths

| Language | Ceiling | Minimum / mean / maximum | Mean use | Distractors |
|----------|---------|--------------------------|----------|-------------|
| English  | 2,048   | 1150 / 1329.7 / 1454     | 64.9%    | 9-14        |
| English  | 4,096   | 2580 / 2708.6 / 2860     | 66.1%    | 21-26       |
| Spanish  | 2,048   | 1766 / 1959.5 / 2044     | 95.7%    | 9-14        |
| Spanish  | 4,096   | 3844 / 4008.5 / 4096     | 97.9%    | 21-26       |

Whole passages are packed until the next distractor would overflow any matched language/position prompt. Thus 2,048 and 4,096 are input-token ceilings, not exact matched lengths. The middle condition uses the middle of passage order, not an exact token midpoint. Token counts include instructions, question, options and chat overhead; the generated choice token is additional. Language and actual token length vary together in this design, so the language comparisons and language-by-position interactions cannot isolate a language effect from token-length effects.

The frozen data adapter preserves the official release, verifies parallel question IDs and gold labels, and adds `split=official-unspecified` as an identity placeholder. That field does not create a train/test partition. The seed-17 sample is held out from development; synthetic fixtures were used for software and format checks.

## Observed response patterns after scoring

There were 0 invalid responses among 1,600 committed outputs. The table records observed outcome patterns after scoring. Patterns overlap and are not causal error categories. Examples follow a deterministic pair-ID selection rule; they are illustrative, not a representative error sample. No explanation of model reasoning is inferred from a one-letter output. Full question, options, evidence and raw responses are preserved in `observed-errors.json` when a manifest is supplied.

| Language / ceiling | Edges right; middle wrong | Middle right; edges wrong | Gold-only right; middle wrong | All positions wrong |
|--------------------|---------------------------|---------------------------|-------------------------------|---------------------|
| English / 2,048    | 2                         | 1                         | 15                            | 33                  |
| English / 4,096    | 7                         | 0                         | 20                            | 30                  |
| Spanish / 2,048    | 5                         | 3                         | 22                            | 38                  |
| Spanish / 4,096    | 8                         | 0                         | 22                            | 34                  |

Every count has the same denominator: 100 paired questions in the specified language and ceiling. Categories overlap; do not add their counts.

### Observed example 1: English, 2,048 tokens

Pair 4dddf6d28af4ee1b...; pattern both edges correct middle wrong; gold option B.

Question: According to the passage, which of the following is most likely to be problematic for unprepared travelers visiting Greenland?

Raw outputs: First = 'B' (correct); Middle = 'D' (incorrect); Last = 'B' (correct); Gold passage only = 'B' (correct); No context = 'B' (correct).

### Observed example 2: English, 2,048 tokens

Pair 3ddcc6cec650887e...; pattern middle correct both edges wrong; gold option D.

Question: What policy is in place to protect the canyon?

Raw outputs: First = 'B' (incorrect); Middle = 'D' (correct); Last = 'B' (incorrect); Gold passage only = 'D' (correct); No context = 'D' (correct).

## Scope of the evidence

This is one model, one quantization, two languages and a 100-question sample; it does not identify an algorithmic contribution or establish general performance rankings. Translation, tokenization, instruction language, answer priors, related distractors and unknown training exposure may affect results. Intervals describe resampling uncertainty within the selected passage population; they do not cover every source of bias. No prompts, model choices or analysis definitions should be revised using these held-out results. Follow-up hypotheses require a separately frozen study.

## Audit trail

The generator recomputes the entire score from checksum-verified committed prediction records before writing outputs. For the benchmark it also verifies the frozen study seal, source files, protocol, manifest, matrix, model fingerprints and execution environment. Local hashes document file consistency; they are not independent proof of when a researcher first viewed the data. Observed runtime includes compilation and machine contention and is not a controlled hardware benchmark.

Run ID: eb98eea93d608a94bec3cf0342a65fbcaeecb1a409f7a027108b216f81ef1a8

Matrix ID: 646a2c8f9d51c41cad0ce8edc2c67eea7dd47e78a8ca9348c11e017a8081a68

Study ID: abec75903504552bc15670b61bd3101ab8d63d52766f3036f3f1d256e0cf79c7. Frozen at 2026-10-09T16:48:57.439977+00:00.

Model: mlx-community/Qwen3-1.7B-4bit. Revision: 3b1b1768f8f8cf8351c712464f906e86c2b8269e.

## Reproduction and supporting files

report.md includes every accuracy and paired-contrast estimate. CSV tables provide machine-readable values; observed-errors.json contains selected questions and raw responses; provenance.json binds all generated files. report-generator.py is the exact script used. The original sealed study, matrix, score and prediction journal remain the authoritative inputs.

## Sources and attribution

Liu et al. (2023). Lost in the Middle: How Language Models Use Long Contexts. <https://arxiv.org/abs/2307.03172>

Belebele benchmark and data license (CC-BY-SA-4.0): <https://github.com/facebookresearch/belebele>. Dataset text reproduced in examples remains subject to its source attribution and license.

MLX-LM local inference: <https://github.com/ml-explore/mlx-lm>. Model source: <https://huggingface.co/mlx-community/Qwen3-1.7B-4bit>