Multiplication in Multimodal LLMs

Computation with text, image, and audio inputs

Findings: Modality changes how problem difficulty predicts accuracy, how model internals operate.

Given the same problem on different modalities, model outputs differ

Arithmetic load can flip correctness

40,000 × 30,000
sparse 5-digit operands

$C = 10 \times 2 = 20$

45,327 × 36,215
dense 5-digit operands

$C = 10 \times 10 = 100$

Same broad task,
but difference in load placed on model.

Why the multimodal need

- Agents inspect screenshots, slides, scans, forms, and voice traces.
- Benchmarks often mix perception, planning, and computation.
- Exact arithmetic requires a hard answer, not a plausible explanation.

therefore we need paired tasks where only the modality, surface form changes

BENCHMARK

Hold the arithmetic fixed; vary the modality, surface form

generator → paired renderings → same problem different form → model, modality comparison

Same instance, three channels

- **text** 47 × 36 = ?
- **image** rendered equation
- **audio** spoken number words

What the design buys us

10,000 shared instances

C digit load

5 modalities /surface forms

Paired instances isolate arithmetic load from perception and wording.

hf.co/datasets/cjerzak/MultimodalMathBenchmarks

A simple arithmetic load variable to quantify difficulty

C = total digits × non-zero digits

A simple proxy for how many digit interactions the model must carry.

C is a simple proxy; the paper also compares it to more exact intermediate-step counts.

Example: 47 × 36

4 7 × 3 6

4 total digits × 4 non-zero digits = C 16

Problem structure changes computational complexity, in addition to number size.

9 9 9 9 × 8 8 8 8

C = 8 × 8 = 64

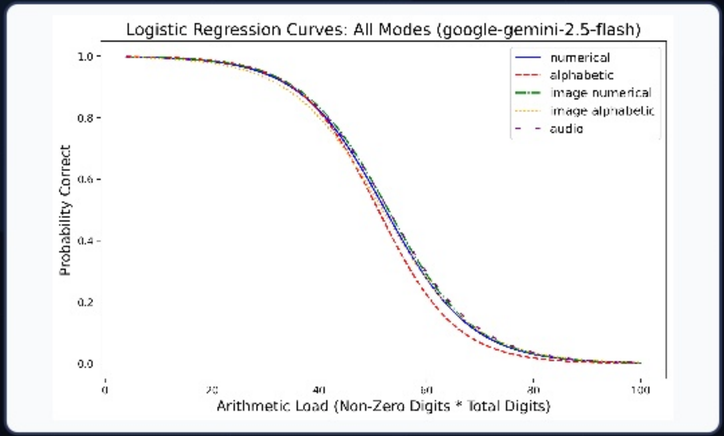
9 0 0 0 × 8 0 0 0

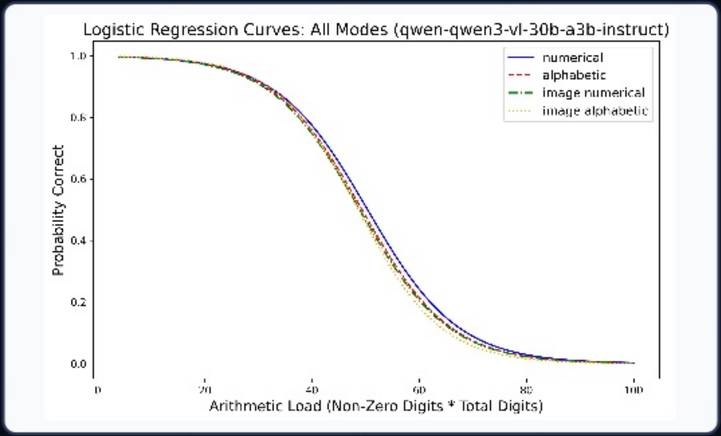
C = 8 × 2 = 16

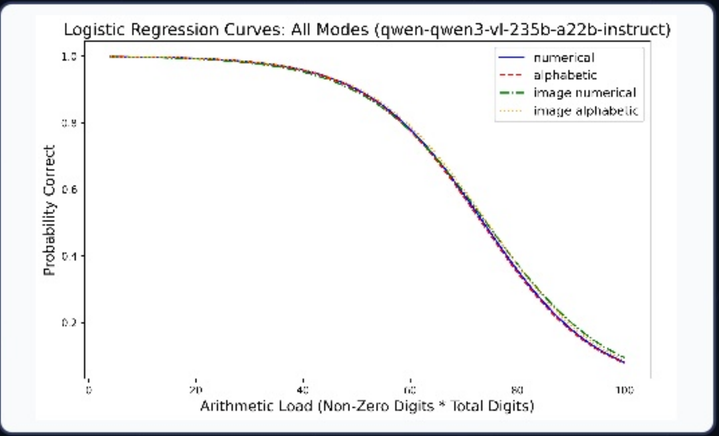
FINDING

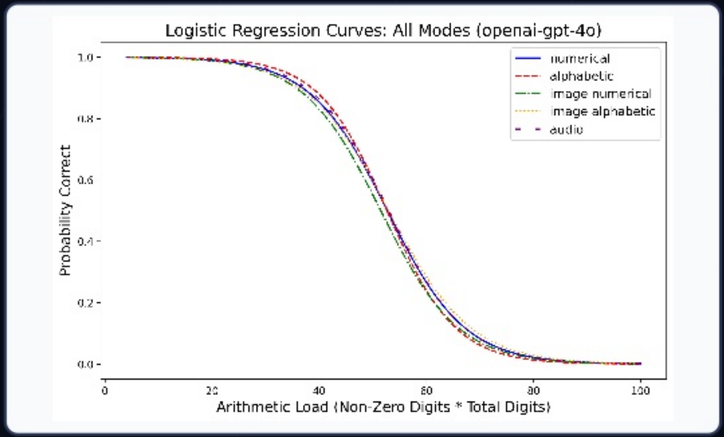
Result 1: accuracy falls off a load cliff

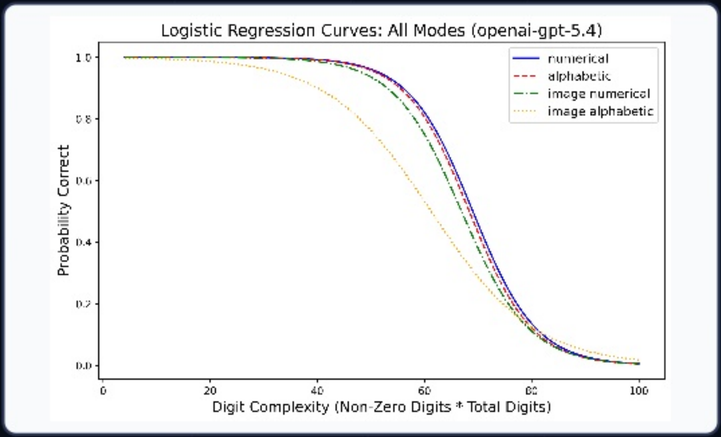
Logistic-regression panels: probability correct versus arithmetic load C across modes; see paper for R-squareds

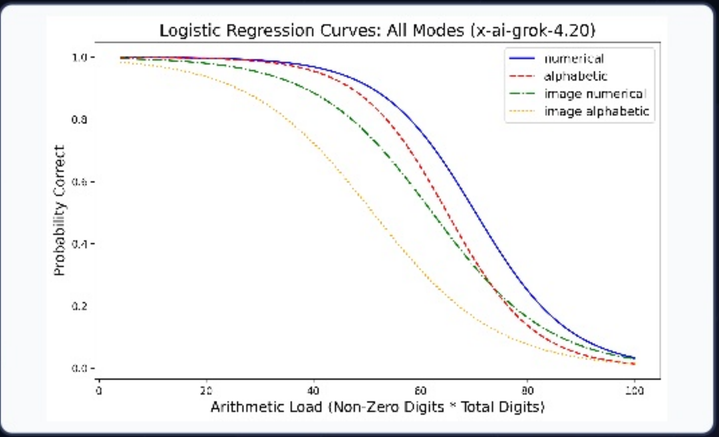








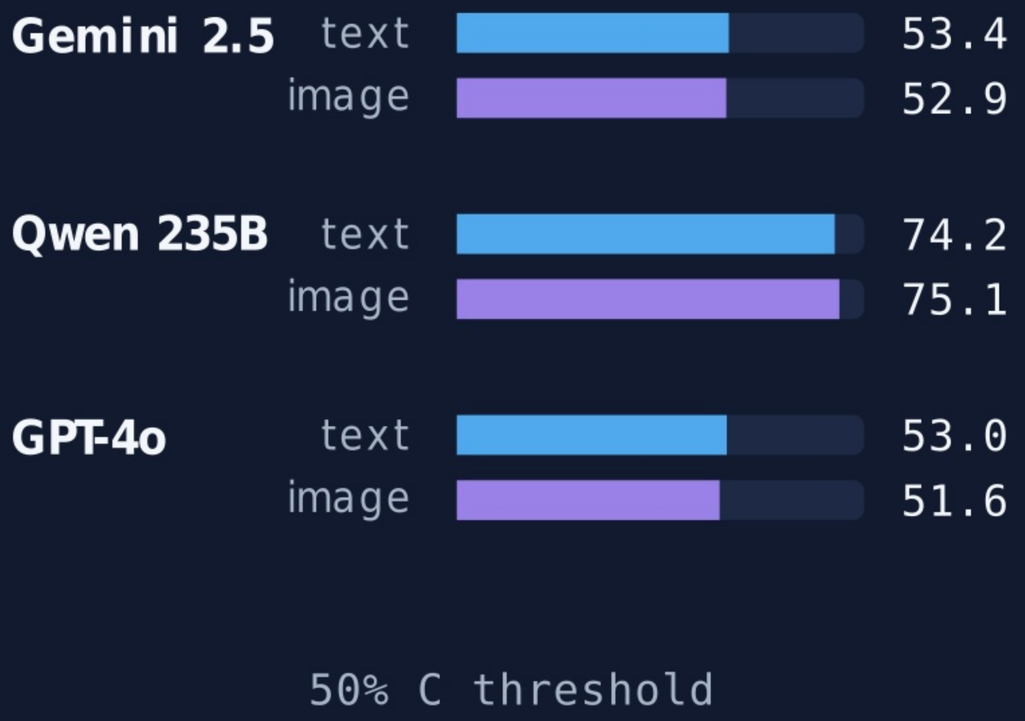




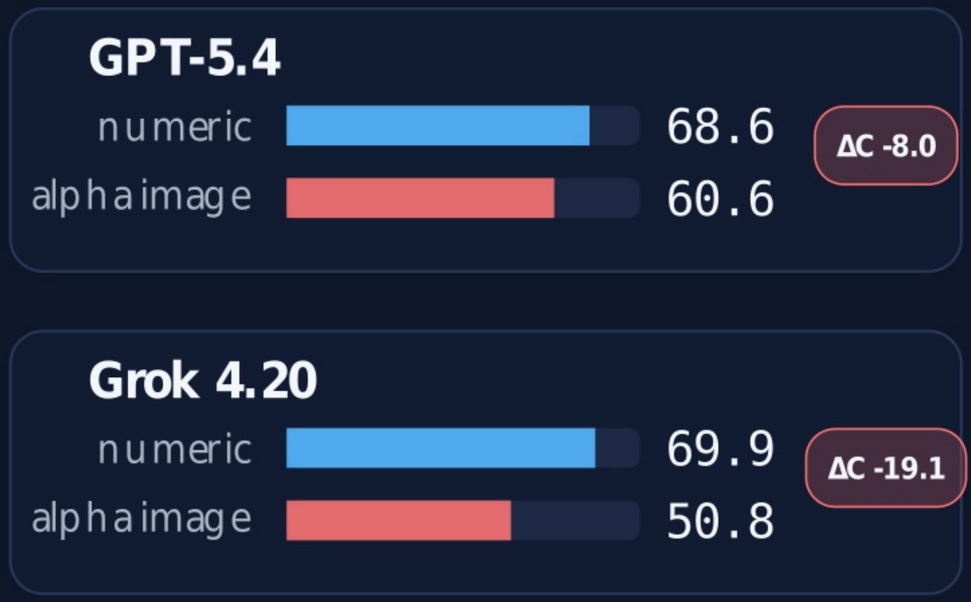
Arithmetic load is the dominant degradation axis; modality mainly shifts where each model begins to fail.

But modality changes where the performance cliff begins

Numerical images often track text



Alphabetic images are more fragile



ΔC = alphabetic-image 50% threshold − numeric 50% threshold; negative means the model fails earlier.

DIAGNOSIS

But perception is not the bottleneck

Perception

>99%

matched perception checks across modalities

Computation

C grows

partial products, carries, and long-range digit interactions accumulate

Answer

accuracy drops as C grows.

Modality differences not due to perception errors; they appear after the problem is read.

Quantify internal model routing across modalities to explore mechanism.

OT

Columnar

digit-by-digit
partial products

```
  47
 x36
----
 282
1410
----
1692
```

DD

Decompose

split into place-value
chunks

$$47 \times 36 = 40 \times 36 + 7 \times 36$$

RC

Round + compensate

use nearby round base
then adjust

$$49 \times 51 = 50^2 - 1^2$$

forced-completion loss probe

score heuristic preambles
under the same problem context

lower loss = deeper alignment

FINDING

Result 2: decomposition heuristic is preferred in Qwen models

Lowest forced-completion loss by condition

30B / text

★ **DD: -0.506**
RC: -0.444
OT: -0.440

DD wins

30B / image

★ **DD: +0.916**
OT: +0.978
RC: +1.652

DD wins

235B / text

★ **DD: +1.473**
RC: +1.879
OT: +2.341

DD wins

235B / image

★ **DD: +1.459**
RC: +2.211
OT: +2.272

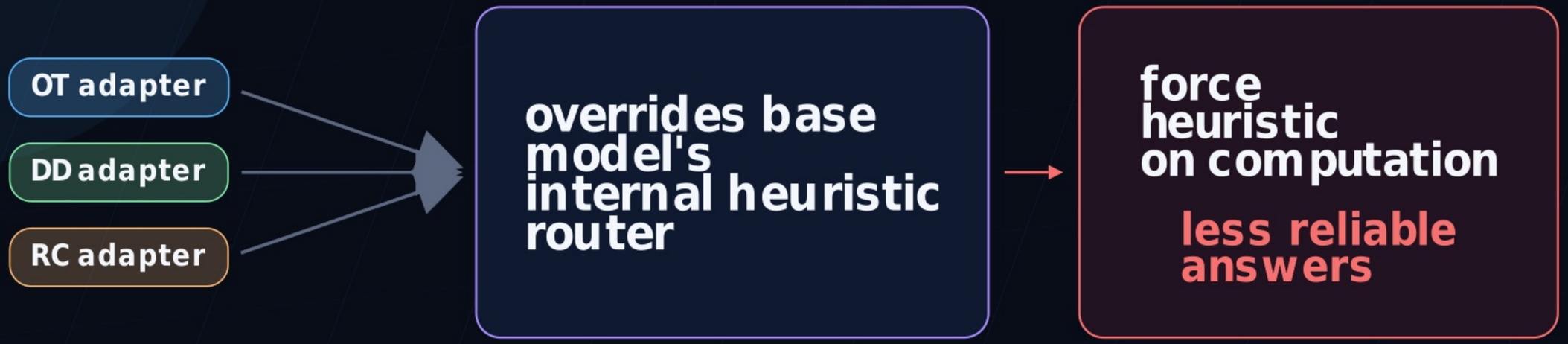
DD wins

Interpretation

- Probe scores continuation fit
- DD lowest loss in all conditions
- Images shift margins not the winner

Accuracy tells us when models fail; loss probes ask what procedure they lean toward before failing.

However, steering towards a single heuristic can make arithmetic worse



Qwen3-VL 30B

base accuracy 74.2%

1	113	114
improved flips	degraded flips	total correctness flips

Qwen3-VL 235B

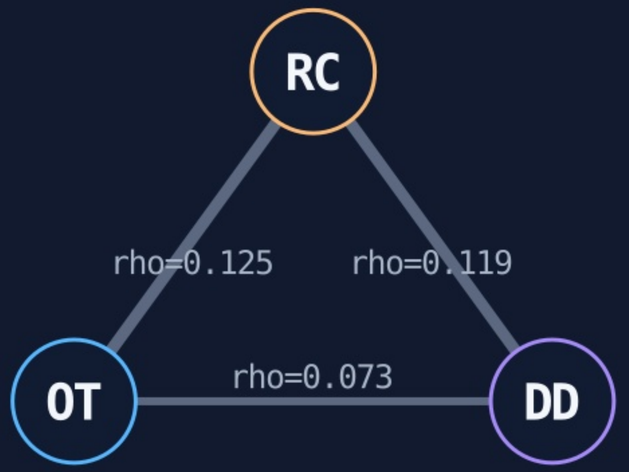
base accuracy 74.8%

4	117	121
improved flips	degraded flips	total correctness flips

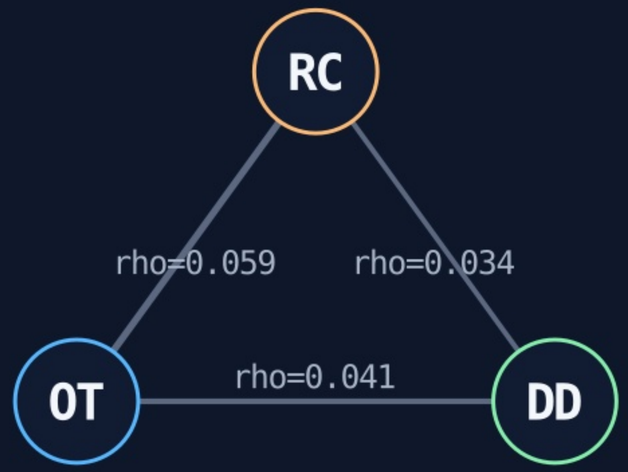
The adapter learns a style, but the base router balances shortcuts better than any single imposed trace.

Why does forced routing degrade performance? Different heuristics contain unique information.

line width ∝ correlation



30B effective LoRA updates



235B effective LoRA updates

Robustness test: Same-heuristic reruns are more similar than cross-heuristic pairs; separation is not high-dimensional noise.

What to remember

1. **Load predicts failure**
   Digit structure drives the accuracy cliff more consistently than modality labels alone.

2. **Seeing is not computing**
   Perception checks are near-perfect; multiplication still collapses as C grows.

3. **Strategy priors are measurable**
   Forced-completion loss reveals a decomposition groove across Qwen text and image inputs.

4. **Internal router beat forced heuristics**
   LoRA nudges change behavior, but often degrade correctness; exact arithmetic needs verification.

For more information, see neuristemic.ai