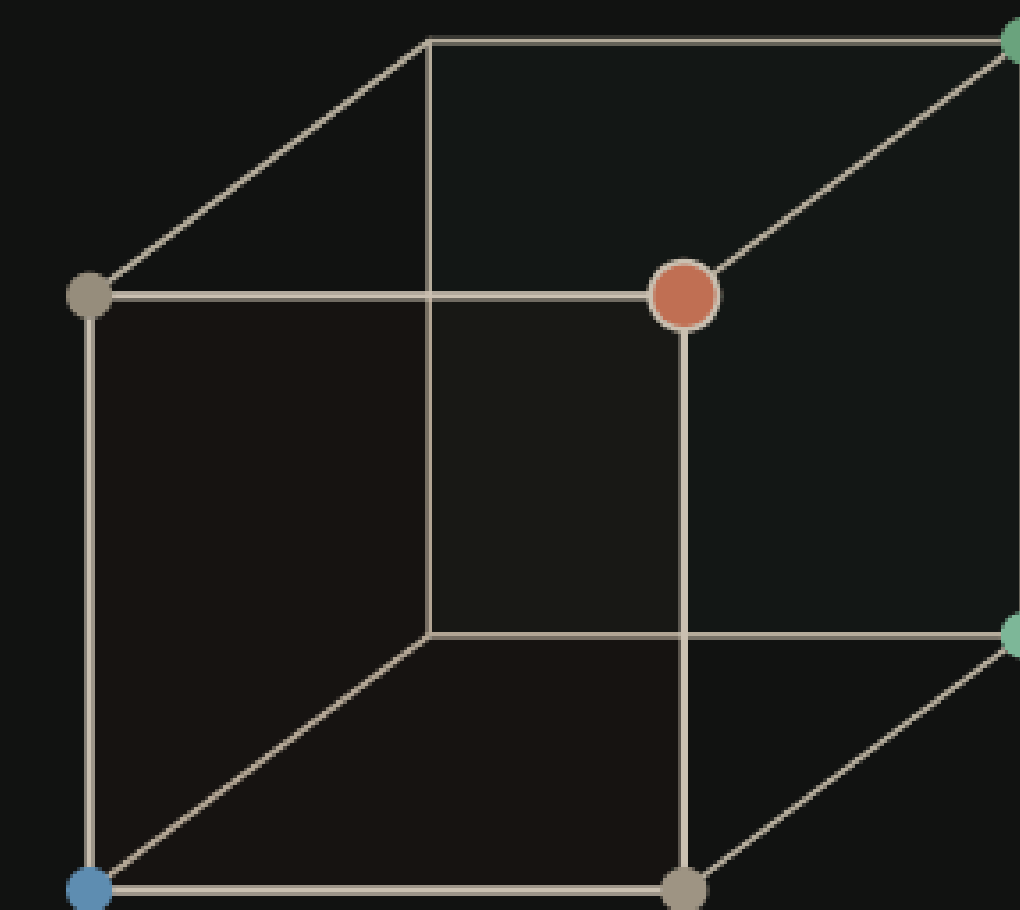


COST-AWARE ROUTING · WEB-USAGE AGENTS

Which way of seeing a web page is worth paying for?

A web agent can read a page as a **screenshot**,
an **accessibility tree**, or a **Set-of-Marks** overlay.
We ask whether **routing** between
representations beats any single one.



Jiaming Wei · MSc dissertation · solo

MSc Artificial Intelligence for Sustainable Development, UCL

~120K LOC Python

4 months

1,141 tests

THE PHANTOM ROUTING SPACE

Six observation modes are the corners of one 2x2x2 cube.

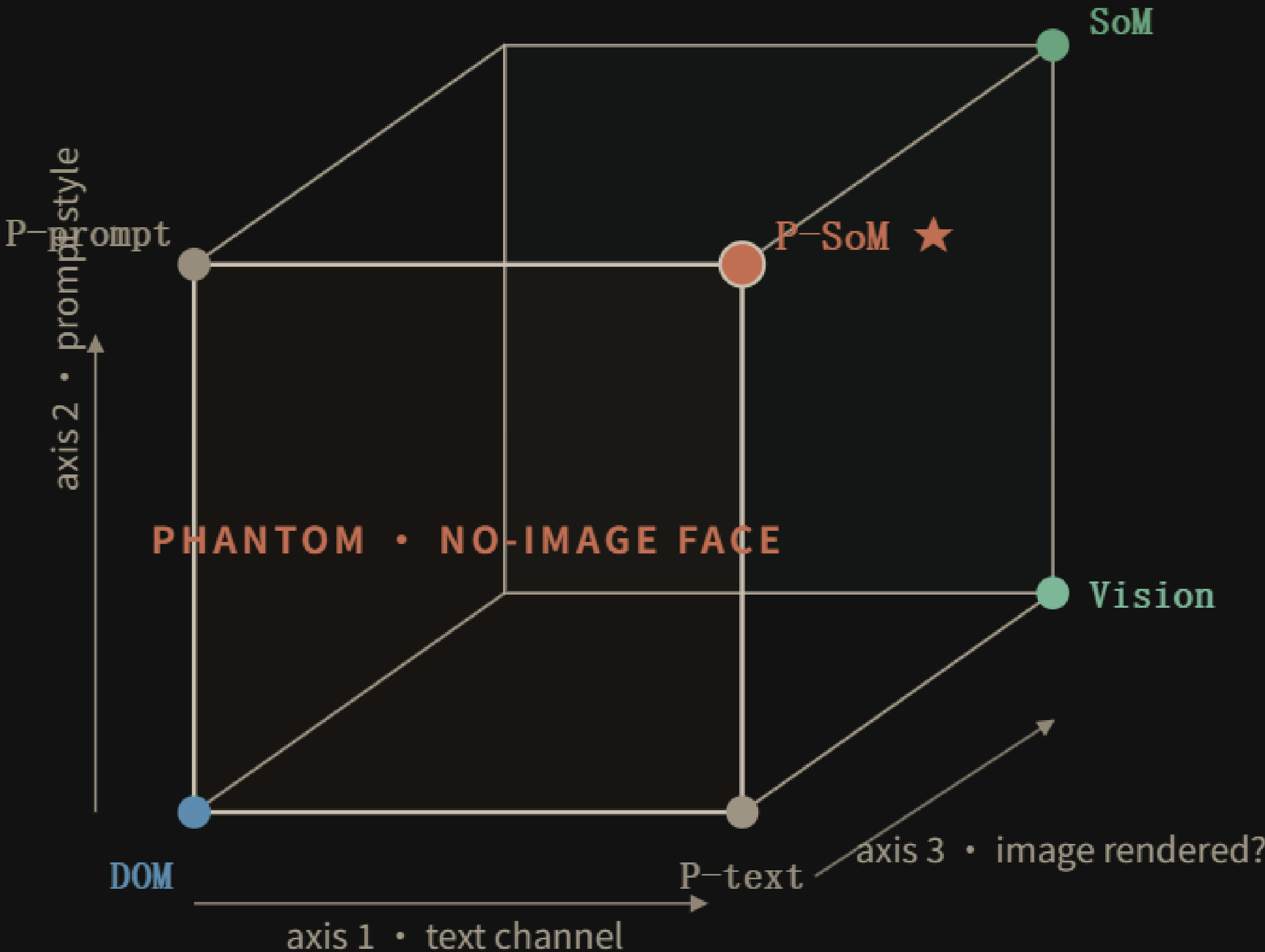
Each mode is a choice on three axes: the **text channel**, the **prompt style**, and **whether a screenshot is rendered** at all.

THE PHANTOM SPACE

The entire no-image face of the cube — four modes that never see a pixel. DOM P-text

P-prompt

P-SoM



EXPERIMENT SETUP

A preregistered, power-analysed sweep of the full cube.

3 MODEL FAMILIES × 6 MODES × 3 VWA SITES

MODEL FAMILY	TIER	MODES
Qwen3-VL-235B	frontier · hosted	6
Qwen3-VL-4B	small · local	6
Gemma-3-4B	small · local	6

Sites · **classifieds** · Reddit · Shopping

Preregistered

Hypotheses, gate threshold and stopping rule fixed *before* the clean run — no post-hoc target-shopping.

Power-analysed

Sample sizes chosen to detect a +1.0pp effect at the target power before committing compute.

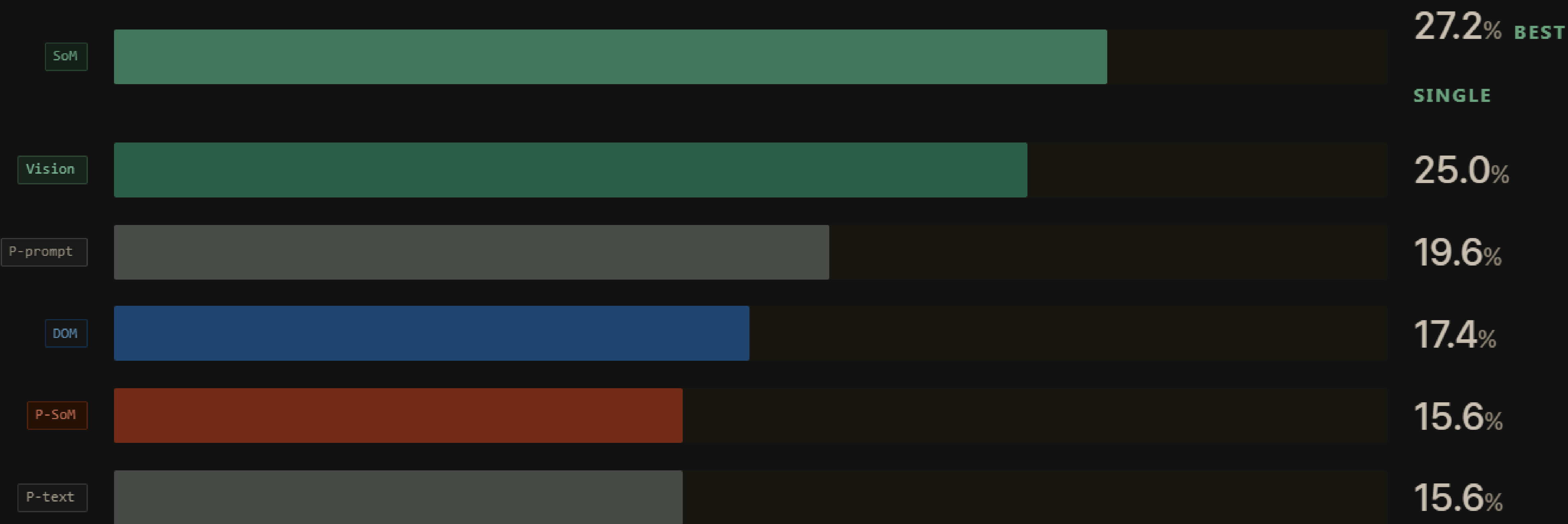
3-tier compute fleet

Frontier hosted API + two local small-model tiers, orchestrated as one race-safe pipeline.

SIX MODES, RAW SUCCESS RATE

No single phantom mode wins on its own.

Task success rate on `classifieds` · Qwen3-VL-235B. The image-on modes lead — every phantom mode sits below them.



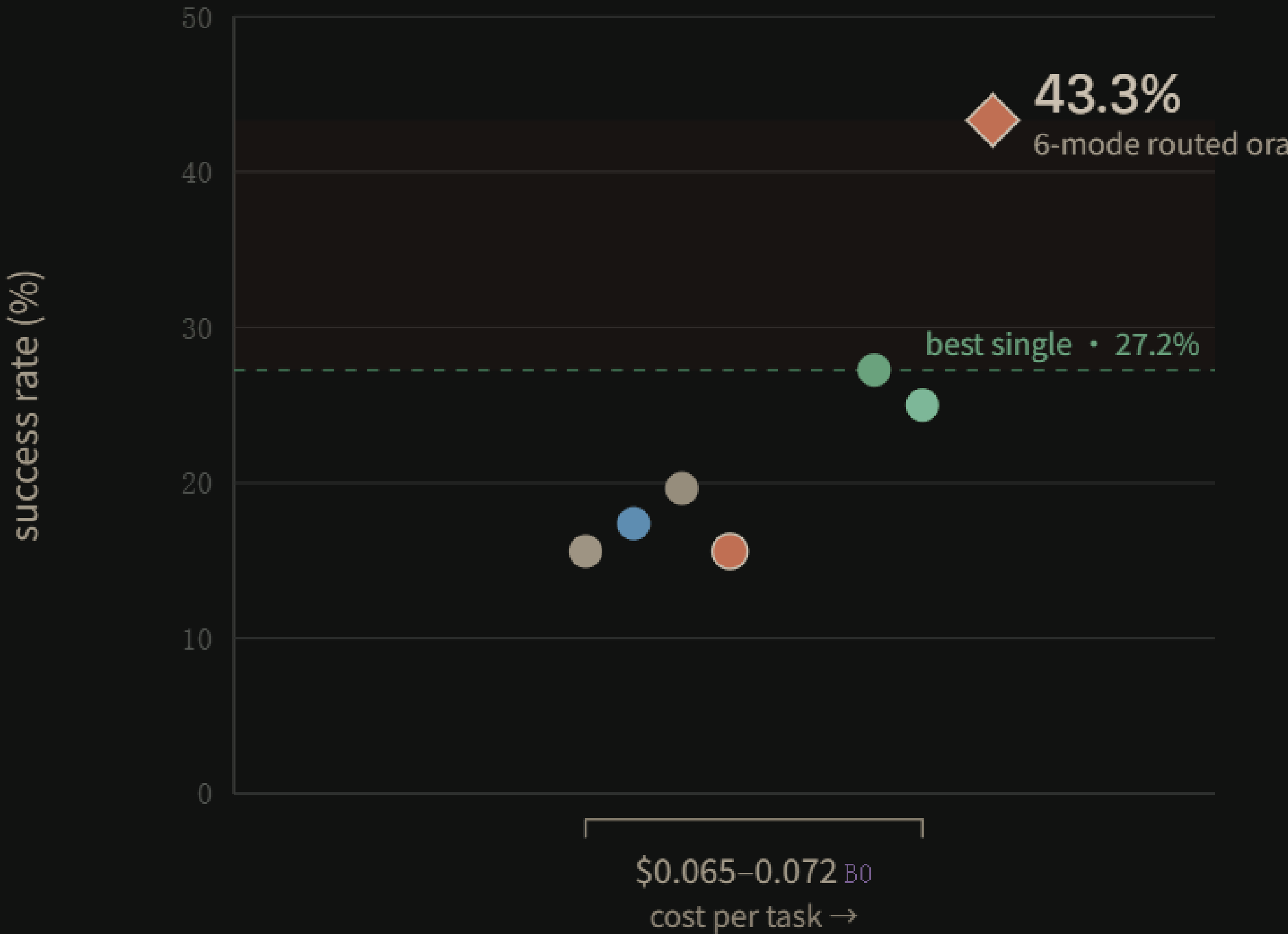
On a flat leaderboard the phantom modes look like the losers. The next two slides show why that reading is wrong.

COST IS THE FLOOR, ROUTING IS THE LEVER

All six modes cost about the same — so spend the budget on **routing**.

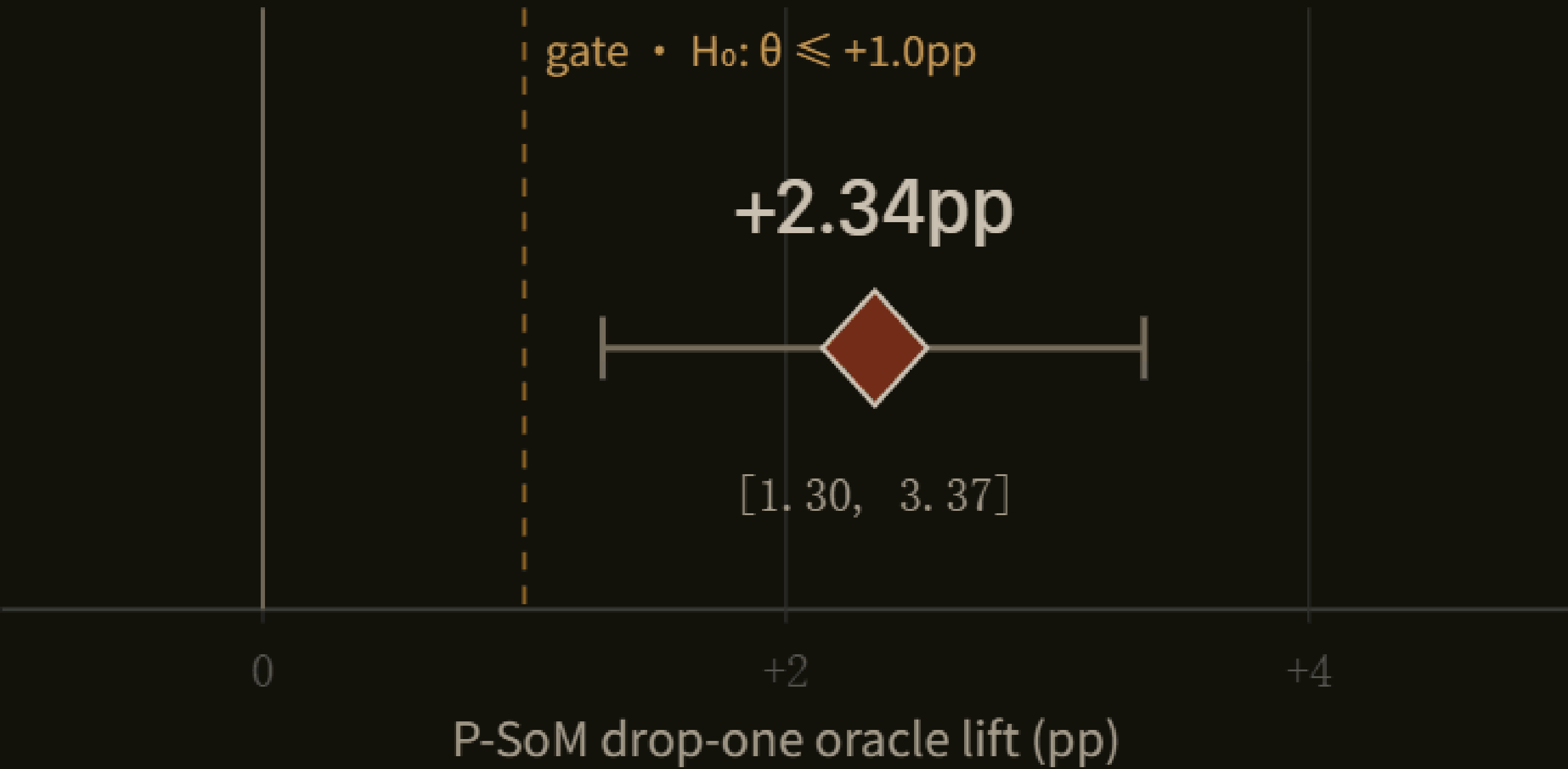
Every mode lands in a narrow **\$0.065–0.072** band per task. Cost can't separate them. But a 6-mode routed oracle — picking the right mode per task — reaches **43.3%**.

43.3% **+16.1_{pp}**
6-mode routed oracle over best single mode



POOLED META-ANALYSIS · K = 3 CELLS

Adding the **P-SoM** arm clears the preregistered bar.



0%
heterogeneity · I^2

p=.0006
vs. preregistered gate

Statistical power **81%** at k=3 today, projected to **97%** when the clean run reaches **k=6** cells.

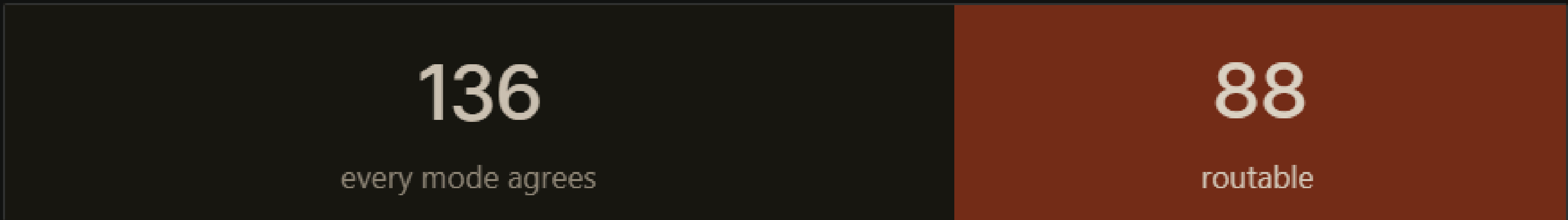
THE HEADLINE

A representation that loses every head-to-head is, statistically, the one most worth routing to.

FAILURE ANALYSIS

Only 88 of 224 tasks are **routable** — that is the whole opportunity.

224 TASKS · PARTITIONED BY ROUTING SIGNAL



No mode helps — fixed ceiling / fixed floor

39% · the entire routing budget

P-SoM rescues image-perception failures

A deterministic failure taxonomy over the 88 routable tasks shows the phantom arm specifically recovers cases where **image perception**, not reasoning, was the bottleneck.

Taxonomy is rule-based and reproducible — no LLM judge in the labelling loop.

ENGINEERING

Every number is guarded by infrastructure.

3

Race-safe orchestration

One distributed scheduler across three compute tiers; no lost or double-counted runs under contention.

6

Contamination watchdog

Six independent layers catch train/test leakage and cross-run bleed before any cell is admitted.

JSONL

Schema-versioned logs

Every result row carries a schema version, so analysis never reads stale shapes.

1,141

Tests on the pipeline

The analysis path that produces these figures is itself covered by 1,141 tests.

Paper-grade discipline: the claims are only as trustworthy as the code that produced them — so the code is tested like the claims matter.

HONEST STATUS

Every number above is real and scope-tagged.

WHAT IS COMPLETE

On `classifieds` all six modes have finished on the largest model (Qwen3-VL-235B). The head-to-head, cost band and pooled lift you just saw are computed on real, completed runs.

WHAT IS STILL FIRING

Other cells of the cube are still running. A clean **k=6** rerun is in progress, and the failure taxonomy is **provisional** until it lands.

No invented numbers

Pending cells render as placeholders, never as zeros and never hidden.

Cost basis disambiguated

Token-billed `B0` costs are never compared against electricity-derived ones.

Power, not just p

The k=6 projection is stated as projected, with its current 81% power shown.

COST-AWARE ROUTING · WEB-USAGE AGENTS

A losing representation can be the most valuable one to route to.

github.com / **Quarkgluonmixture** / Cost-Aware-Routing-for-Web-Usage-Agents

Built on **VisualWebArena**

Jiaming Wei · ucab352@ucl.ac.uk

linkedin.com/in/jiaming-wei-810ab938b

MSc dissertation · solo · ~120K LOC Python · 4 months · 1,141 tests.

MSc Artificial Intelligence for Sustainable Development, UCL