DOES PROBE CONDITIONING ISOLATE COMPOSITION? AN INFORMATION-CONTROL AUDIT OF MULTI-HOP EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditioning multi-hop evaluation on successful component probes is intended to reduce missing-prerequisite confounding. A FULL-FACT condition that supplies both component facts, however, also reveals the answer-bearing second hop, so its performance gain cannot identify composition by itself. We audit this contrast with an information-control suite in MEDCOMPBENCH, an ontology-controlled biomedical evaluation containing 18,856 real chains and 4,711 counterfactual twins. On Qwen2.5-7B, providing only the bridge improves multiple-choice accuracy by .097 [.059,.137], whereas providing only the answer-bearing second fact improves it by .338 [.286,.392] and providing both facts improves it by .389 [.335,.443]. Across seven Qwen2.5 and Qwen3 checkpoints, the full-fact support gap ranges from .259 to .392 for control-passing models. Independent endpoint checks with Phi-3.5-mini and BioMistral-7B yield gaps of .317 [.264,.374] and .202 [.145,.257]. On matched items, neither endpoint shows a resolved difference from Qwen2.5-7B. Across five prespecified new composed-question wordings, the gap remains positive for all 15 checkpoint–wording pairs. Separately, independently reworded component probes retain only 40.7–58.7% of originally admitted items, showing that the conditional population is strongly prompt-dependent. The original contrast is therefore mainly an answer-support effect rather than an isolated measure of bridge retrieval or composition. Conditional multi-hop evaluations should separate bridge provision, answer-bearing facts, and gate stability before assigning a compositional interpretation.

1 INTRODUCTION

Large language models are increasingly evaluated on questions whose answers require several component facts. Benchmark reports often count a correct response as evidence of composition and an error as a reasoning failure, but aggregate accuracy does not distinguish those cases. A model asked for the inheritance pattern associated with a gene may lack the gene–disease association, lack the disease–inheritance association, or fail elsewhere in the multi-hop query. The same ambiguity arises whenever a task's prerequisites can be queried separately.

The issue extends beyond medicine. Multi-hop QA benchmarks can contain shortcuts that make nominally compositional questions solvable from one evidence fragment (Min et al., 2019); controlled construction from single-hop questions reduces but does not eliminate the problem (Trivedi et al., 2022). At the other extreme, an item can be labelled a reasoning failure when the model never had access to the necessary facts. A capability claim is valid only if the measurement isolates its intended construct (Bean et al., 2025). Prior work relates multi-hop and single-hop performance through a compositionality gap (Press et al., 2023) and filters shortcuts before testing latent multi-hop reasoning (Yang et al., 2025b). We complement those approaches by conditioning the unit of analysis on the evaluated model demonstrating both component facts.

We therefore treat demonstrated prerequisite retrieval as part of the evaluation population rather than as an unobserved nuisance. MEDCOMPBENCH turns this idea into an auditable protocol. Each ontology chain $h \to b \to t$ yields two prerequisite probes and three paired composed branches. An item is admitted for model m only after repeated multiple-choice and free-generation success on

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.

both relations. On this conditional set, LATENT poses the closed-book query, COT adds a step-by-step instruction, and FULL-FACT states both component facts. Our measured contrast is

$$\Delta_{\text{full}}(m) = \text{CCA}_{\text{FULL-FACT}}(m) - \text{CCA}_{\text{LATENT}}(m), \tag{1}$$

the paired benefit of full factual support. Because FULL-FACT includes the answer-bearing second-hop fact, Eq. 1 combines bridge availability, second-hop retrieval, contextual reading, and answer extraction. The gate narrows the population; the contrast still has to be decomposed before it can be interpreted as composition.

The intervention results sharpen that interpretation. On the Qwen2.5-7B conditional set, supplying only the bridge improves accuracy by .097 [.059,.137], while supplying the answer-bearing second fact improves it by .338 [.286,.392] and supplying both facts by .389 [.335,.443]. A full-length irrelevant context changes accuracy by $-.047$ [$-.089$,$-.006$], and requiring an explicit bridge line changes it by $-.008$ [$-.033$,.019]. Phi-3.5-mini and BioMistral-7B also show positive full-fact gaps, although bridge-only prompting lowers accuracy for both. Newly worded prerequisite probes retain 40.7–58.7% of the original conditional sets. The gate therefore captures prompt-conditioned access to the component relations, not a stable guarantee that the sampled path will be recovered.

Our contributions are:

1. **Information controls for conditional multi-hop evaluation.** We separate bridge, first-fact, second-fact, full-fact, forced-chain, misleading-fact, and length-control effects instead of assigning the full-fact contrast to composition.
2. **Path-, prompt-, and dependency-aware audits.** We compare licensed, bridge-exact, and edge-exact gates; introduce an independent held-out probe bank; report head- and bridge-cluster uncertainty; expose free-response results; and directly estimate matched-model gap differences.
3. **A reproducible ontology-derived resource.** The frozen bank contains 18,856 real chains and 4,711 counterfactual twins over four HPO/Mondo chain specifications, with deterministic construction, local-neighbourhood distractors, aliases, and explicit set-valued answer sets.

2 RELATED WORK

Factual and compositional probing. Factual probing asks whether parameters encode relations recoverable from prompts (Petroni et al., 2019). Such probes are necessary but do not establish that separately recoverable facts can be composed. Press et al. (2023) formalize a compositionality gap using single- and multi-hop question answering and find that single-hop gains can outpace multi-hop gains. Yang et al. (2025b) show that latent composability depends on relation and shortcut filtering. Concurrent work studies how models compose a broad collection of explicit functions and distinguishes direct from intermediate-variable mechanisms (Khandelwal & Pavlick, 2025). Our focus is measurement: repeated dual-format gates define the evaluation population, information controls separate what the prompt supplies, and named common subsets prevent model and item effects from being conflated.

Multi-hop QA and benchmark shortcuts. HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022) made multi-hop reasoning measurable at scale. Their evidence-grounded setting differs from our closed-book setting, but their central design concern is shared: the intended path must be necessary. Single-hop baselines and ablations show that compositional wording alone does not guarantee multi-hop inference (Min et al., 2019). NeurIPS benchmark papers increasingly treat the evaluation design itself as the scientific object: BABILong controls how distributed facts must be combined (Kuratov et al., 2024), GSM1k uses a matched fresh benchmark to separate generalization from contamination (Zhang et al., 2024), and KGQAGen combines structured grounding with symbolic verification (Zhang et al., 2025). MEDCOMPBENCH follows the same audit-first logic, using ontologies both to generate chains and to verify answers. Context-faithfulness evaluations further show that models can over-rely on supplied false or counterfactual evidence (Ming et al., 2025; Huang et al., 2025). Our wrong-bridge and wrong-second-hop branches use that sensitivity as manipulation checks: a response to supplied facts is not by itself evidence that the intended latent path was composed.

Biomedical QA. Medical benchmarks such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and MultiMedQA (Singhal et al., 2023) measure broad clinical knowledge and answer quality. MEDCOMPBENCH is not a clinical decision benchmark: no patient records or treatment recommendations are involved. Medicine supplies a structured and versioned relational substrate through HPO (Köhler et al., 2021) and Mondo (Vasilevsky et al., 2022). This permits exact chain construction, ontology-native aliases, and explicit accounting for set-valued relations.

3 MEDCOMPBENCH

3.1 CONDITIONAL EVALUATION POPULATION

Each real item i consists of two asserted relations, $r_A(h, b)$ and $r_B(b, t)$, whose composition licenses an answer t. Five frozen paraphrases probe each relation in multiple-choice form and the same five paraphrases are rendered without answer options. Let $P^j_{m,i,k}$ and $F^j_{m,i,k}$ denote multiple-choice and free-response correctness for hop $j \in \{A, B\}$ and paraphrase $k \in \{1, \ldots, 5\}$. The implemented gate is

$$G_m(i) = \prod_{j \in \{A,B\}} \mathbb{1}\left[\sum_k P^j_{m,i,k} \geq 4\right] \mathbb{1}\left[\max_k F^j_{m,i,k} = 1\right]. \tag{2}$$

The model-specific conditional set is $S_m = \{i : G_m(i) = 1\}$. CCA is the multiple-choice accuracy of a composed branch on S_m. Free generation prevents admission based only on option recognition, but the any-of-five rule is permissive and sensitive to phrasing. Grading normalizes punctuation and case, accepts ontology synonyms, and, for one-to-many relations, accepts every ontology-licensed tail.

The gate changes both interpretation and denominator. It does not estimate full-bank two-hop accuracy; it describes behaviour among items whose broad component relations were produced under this prompt bank. We always report $n_{\text{cond}} = |S_m|$ and coverage $|S_m|/18{,}856$. The frozen v3 runs did not evaluate composed branches on every bank item, so an unconditional-vs-conditional comparison cannot be reconstructed and is not presented as a result.

Set-valued grading creates a path-specificity distinction. The *licensed* gate in Eq. 2 accepts any valid relation tail. Our post-hoc *bridge-exact* gate additionally requires the item bridge b on hop A; the *edge-exact* gate also requires the sampled tail t on hop B, under the same $4/5$ MCQ and any-of-five free rule. The item bank itself contains multiple legal hop-A tails for 10,978/11,643 disease–phenotype–phenotype-class items. Only 315/18,856 generated hop-A MCQs expose more than one valid option, but free responses can still validate a different bridge. These exact gates are sensitivity estimands, not changes to the preregistered result.

Let $Y^z_{m,i} \in \{0, 1\}$ denote correctness for composed branch $z \in \{\text{LATENT}, \text{COT}, \text{FULL-FACT}\}$. The finite-bank contrast can be written as

$$\widehat{\Delta_{\text{full}}}(m) = \frac{1}{|S_m|} \sum_{i \in S_m} \left(Y^{\text{FULL-FACT}}_{m,i} - Y^{\text{LATENT}}_{m,i}\right). \tag{3}$$

Equation 3 is the *full-fact support gap*. It is a paired behavioural contrast, not a composition estimand: FULL-FACT supplies $h \to b$ and $b \to t$, so it changes bridge availability, second-hop retrieval, contextual reading, and answer extraction together. The .80 full-fact control below only checks that the supported endpoint is reliably solvable. Probe success is evidence of prompt-conditioned accessibility, not proof of a stable internal belief.

The distinction is visible from a simple decomposition. Let $\pi_m = |S_m|/N$ be gate coverage over a fixed bank of N items, and let A^{LATENT}_m be unconditional latent accuracy. Then

$$A^{\text{LATENT}}_m = \pi_m \, \text{CCA}_{\text{LATENT}}(m) + (1 - \pi_m) \, \Pr\left(Y^{\text{LATENT}}_{m,i} = 1 \mid G_m(i) = 0\right). \tag{4}$$

An aggregate score therefore mixes gate coverage (π_m), conditional closed-book performance ($\text{CCA}_{\text{LATENT}}$), and performance where at least one prerequisite was not demonstrated. Conditioning identifies the second quantity for each model, but it also makes S_m model-specific. Cross-model comparisons consequently require a named intersection rather than silently comparing different selected populations.

Table 1: Ontology-derived real-chain specifications. "Set" means that every ontology-licensed tail is accepted. The bank additionally contains 4,711 counterfactual twins.

Composition	Real	Answer
Gene→disease→inheritance	3,207	unique
Disease→phenotype→class	11,643	set
Gene→disease→disease class	595	unique
Gene→disease→phenotype	3,411	set
Total	18,856	

3.2 Composed branches and identification controls

All three branches use the same item and four answer options. LATENT asks directly for t from h and instructs the model to give only the answer. CoT poses the identical relation but requests step-by-step reasoning and an exact `Final answer:` line. FULL-FACT prepends both asserted facts, including the answer-bearing $b \to t$ statement, and asks for t. The artifact key remains `ORACLE` for backward compatibility, but the paper uses FULL-FACT to prevent it from being mistaken for a bridge-only intervention.

The post-audit identification suite holds the question, option order, and decoder fixed while varying supplied information. BRIDGE-ONLY states only the intermediate entity b; FACT-1 states $h \to b$; FACT-2 states $b \to t$; and FORCED-CHAIN requires separate `Bridge:` and `Final answer:` lines. The wrong-bridge branch substitutes a same-relation donor bridge together with its second-hop fact, and the wrong-second-hop branch replaces only the tail. LENGTH-CONTROL supplies an unrelated two-fact context whose complete prompt has exactly the same tokenizer length as FULL-FACT. Wrong-bridge estimates use the valid-donor subset and report its denominator. We evaluate both S_m and a deterministic balanced sample of 400 chains per specification. One disjoint holdout bank adds five new MCQ and free-response probes per hop to test gate retention. A second post-hoc bank varies the composed question over five unused wordings while fixing items and option order; it tests endpoint wording sensitivity without redefining admission or the primary estimand.

The full-fact branch also serves as an endpoint control. We mark checkpoints with $\mathrm{CCA}_{\text{FULL-FACT}} < .80$ as *control-failed* and exclude their latent scores from cross-model comparisons while retaining them in all tables and figures. The .80 cutoff was fixed before the v3 comparisons; .75 and .85 appear as sensitivity analyses. A second endpoint,

$$\Delta_{\text{CoT}}(m) = \mathrm{CCA}_{\text{FULL-FACT}}(m) - \mathrm{CCA}_{\text{CoT}}(m), \tag{5}$$

asks how much of the full-fact support advantage remains after an explicit reasoning instruction.

3.3 Ontology construction

The bank is a deterministic function of HPO and Mondo releases dated 2025-05-06, seed 20260824, and fixed per-specification quotas. Table 1 lists the four chain types. The construction rejects invalid compositions, duplicate entities, bridge-as-answer collisions, and items without relation-consistent distractors. Distractors come from the item's ontology neighbourhood and an item-ID-derived random seed, rather than a global candidate pool; increasing a quota therefore leaves all earlier items byte-identical. The real bank is accompanied by 4,711 isomorphic counterfactual twins used as a separate in-context binding ceiling.

The specifications intentionally vary semantic distance. Two relations are set-valued: a disease can have multiple phenotypes, and a gene can license multiple phenotypes through associated diseases. Their answer keys store the full valid tail set. The disease-class chain crosses HPO and Mondo and has only 595 constructible items; it is retained but its gated strata are underpowered. The phenotype-class specification contributes 11,643/18,856 real items and consequently dominates pooled conditional scores. We therefore report both pooled CCA and an equal-spec macro average over specifications with $n_{\text{spec}} \geq 50$.

Table 2: Conditional multiple-choice results under the implemented licensed gate. The gap is FULL-FACT−LATENT; SI denotes its 95% bridge-cluster resampling stability interval. Coverage is relative to 18,856 real items. † marks FULL-FACT accuracy below the fixed .80 control threshold.

Model	n_{cond}	Coverage	LATENT	CoT	FULL-FACT	Gap [95% SI]	Control
Qwen2.5-0.5B	159	0.84%	.403	.434	.453	.050 [-.034, .133]	fail†
Qwen2.5-1.5B	453	2.40%	.647	.651	.974	.327 [.276, .382]	pass
Qwen2.5-3B	829	4.40%	.574	.642	.966	.392 [.340, .445]	pass
Qwen2.5-7B	970	5.14%	.594	.665	.982	.389 [.334, .442]	pass
Qwen3-0.6B	112	0.59%	.589	.536	.848	.259 [.154, .365]	pass
Qwen3-1.7B	311	1.65%	.505	.363	.775	.270 [.158, .391]	fail†
Qwen3-8B	885	4.69%	.646	.678	.988	.341 [.292, .389]	pass

4 EXPERIMENTAL PROTOCOL

Models and decoding. We evaluate Qwen2.5-Instruct at 0.5B, 1.5B, 3B, and 7B parameters (Qwen Team, 2024), and Qwen3 at 0.6B, 1.7B, and 8B (Yang et al., 2025a). Qwen3-4B is missing because the frozen run preserved disk for the priority 8B endpoint; no checkpoint was substituted. The post-audit identification suite repeats the full gate and all information controls for Qwen2.5-7B, Microsoft's independently developed Phi-3.5-mini-instruct (Abdin et al., 2024), and BioMistral-7B, a biomedical Mistral derivative (Labrak et al., 2024). These three checkpoints form a model-lineage robustness check, not a size-scaling series. All runs use the official weights in bfloat16, without quantization, temperature 0, top-p 1, and seed 20260824. Short branches have a 32-token generation limit; CoT has a branch-local 512-token limit so its trace is not truncated. Qwen3's hybrid thinking mode is explicitly disabled with `enable_thinking=False`; three stored outputs per completed checkpoint contain no thinking block.

Two-stage evaluation. Stage 1 evaluates PROBE-A and PROBE-B over all 18,856 real chains and computes S_m with Eq. 2. It issues $18{,}856 \times 2$ hops $\times 5$ paraphrases $\times 2$ formats $= 377{,}120$ requests per complete checkpoint. For Qwen2.5, 39,080 frozen legacy rows plus 338,040 extension rows form the same effective total. Stage 2 evaluates LATENT, COT, and FULL-FACT once in each format only for admitted real items, plus the available counterfactual twins. Thus Qwen3-8B uses $885 \times 3 \times 2 + 217 \times 2 = 5{,}744$ Stage-2 requests. Branch-specific decoding parameters enter the generation cache key, so the longer COT budget does not invalidate other branches.

Statistics. Ontology paths share heads, bridges, and edges, so items are not independent. Equation 3 is exactly measured on the frozen bank and has no sampling error for that bank. The reported bootstrap ranges are therefore *cluster-resampling stability intervals*: they describe sensitivity to reweighting observed bridge clusters, not confidence intervals for a defined population of future ontology chains. Bridge is the primary perturbation because it is the repeated mediator in the chain; head-cluster and item intervals are secondary checks. This single-axis bootstrap does not fully model crossed head–bridge dependence, so we avoid superpopulation inference. The 7B licensed set has 970 items but only 486 unique bridges, with as many as 63 items per bridge. Exact McNemar and Holm values from the frozen analysis are retained as descriptive legacy diagnostics, not treated as cluster-valid inference. Popularity bins are terciles of bridge degree computed on the full bank before gating. Per-specification conclusions require $n \geq 50$; smaller rows are underpowered. Because S_m differs across models, cross-model claims require an explicit intersection. The all-seven intersection is empty; $S_2 = S_{\text{Qwen2.5-7B}} \cap S_{\text{Qwen3-8B}}$ contains 341 identical items.

5 RESULTS

5.1 EXPLICIT FULL FACTS HELP AFTER COMPONENT-PROBE SUCCESS

Table 2 reports the implemented licensed gate. Qwen2.5-0.5B and Qwen3-1.7B fail the .80 full-fact control. At the other five checkpoints, Δ_{full} ranges from .259 to .392. The bridge-cluster stability intervals stay above zero; Qwen2.5-7B gives .389 [.334,.442], and Qwen3-8B gives .341

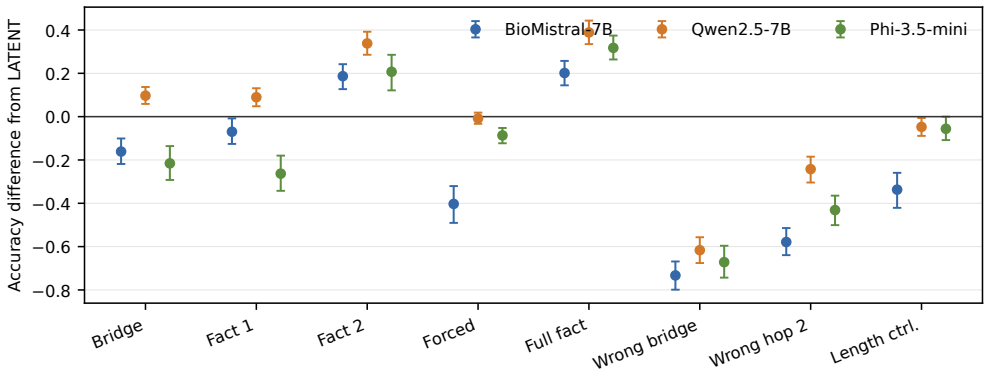


Figure 1: Identification contrasts relative to LATENT on each endpoint model's licensed conditional set. Points show paired accuracy differences; bars are bridge-cluster 95% stability intervals. Wrong-bridge points use the valid-donor subset and report their denominator in Appendix Table 6. FULL-FACT is the frozen artifact branch named ORACLE.

Table 3: Direct full-fact support cases: both licensed component probes pass, FULL-FACT is correct, and LATENT is wrong.

Model	Count/n_{cond}	Proportion
Qwen2.5-0.5B	27/159	.170
Qwen2.5-1.5B	151/453	.333
Qwen2.5-3B	333/829	.402
Qwen2.5-7B	377/970	.389
Qwen3-0.6B	34/112	.304
Qwen3-1.7B	109/311	.350
Qwen3-8B	309/885	.349

[.292,.389]. Explicit full-chain support therefore produces a large paired gain after broad component-probe success. The following interventions locate the source of that gain.

The same pattern can be stated without a difference score. Table 3 counts items that pass both licensed probes, are correct with both facts in context, and fail closed-book: 151/453, 333/829, and 377/970 for Qwen2.5-1.5B/3B/7B, and 309/885 for Qwen3-8B. These are direct full-fact support cases, not a taxonomy of composition errors.

5.2 The full-fact gap is mainly a second-hop support effect

Figure 1 dissects the information contrast at Qwen2.5-7B. On its licensed conditional set ($n = 970$), BRIDGE-ONLY raises accuracy from .594 to .691, a paired difference of .097 with bridge-cluster stability interval [.059,.137]. FACT-1 gives +.090 [.048,.131]. FACT-2 reaches .932, a gain of .338 [.286,.392], while FULL-FACT reaches .982, a gain of .389 [.335,.443]; the second fact alone recovers 87.0% of the full-fact improvement. The token-for-token FULL-FACT length control lowers accuracy to .546, $-.047$ [$-.089$,$-.006$], excluding prompt length as an explanation for the full-fact gain. On the deterministic balanced-bank sample ($n = 1{,}600$), bridge-only and first-fact prompts decrease accuracy by .105 and .133, while the second fact and full facts increase it by .403 and .474. Conditioning changes both the population and the observed intervention effects.

The directional controls show that the supplied relations affect the answer. On the 438 conditional items with a valid donor, replacing the bridge and its continuation reduces accuracy by .616 [$-.676$,$-.557$]; replacing only the second-hop tail reduces it by .242 [$-.304$,$-.185$] over all 970 items. FORCED-CHAIN changes accuracy by $-.008$ [$-.033$,.019]. It names the sampled bridge on 126/970 items, and only 6 of those responses also give the correct final answer. Together, these results attribute most of the full-fact gain to the answer-bearing second fact, while explicit chain elicitation rarely recovers the sampled path.

Table 4: Path-specificity sensitivity for the matched-size endpoints. The gap is FULL-FACT−LATENT; intervals use a bridge-entity cluster bootstrap. Licensed accepts any ontology-valid tail, bridge-exact requires the item bridge on hop A, and edge-exact additionally requires the sampled hop-B tail.

Model	Gate	n	Gap [95% stability interval]
Qwen2.5-7B	Licensed	970	.389 [.334, .442]
	Bridge-exact	474	.392 [.331, .453]
	Edge-exact	409	.330 [.265, .397]
Qwen3-8B	Licensed	885	.341 [.292, .389]
	Bridge-exact	491	.332 [.281, .383]
	Edge-exact	395	.228 [.167, .289]

Table 5: Post-hoc gate-reliability sensitivity. Admission requires the listed number of correct free-response paraphrases on *each* hop in addition to $\geq 4/5$ MCQ. The composed predictions are unchanged. Intervals are bridge-cluster resampling stability intervals.

Model	Free gate	n	FULL-FACT−LATENT [95% interval]
Qwen2.5-7B	1/5	970	.389 [.334, .442]
	2/5	468	.380 [.307, .458]
	3/5	224	.424 [.364, .481]
	4/5	111	.369 [.275, .449]
	5/5	28	.393 [.226, .571]
Qwen3-8B	1/5	885	.341 [.292, .389]
	2/5	423	.324 [.264, .380]
	3/5	229	.306 [.240, .369]
	4/5	131	.282 [.195, .369]
	5/5	46	.174 [.070, .283]

The independently developed checkpoints show the same answer-support pattern but different responses to bridge hints. Phi-3.5-mini ($n = 589$) gives FULL-FACT−LATENT=.317 [.264,.374] and second-fact-only=.207 [.121,.285], whereas bridge-only is $-.216$ [$-.292$,$-.136$]. BioMistral-7B ($n = 273$) gives .202 [.145,.257], .187 [.127,.242], and $-.161$ [$-.219$,$-.101$], respectively. On pairwise matched sets, Qwen-minus-Phi and Qwen-minus-Bio full-fact-gap differences are .039 [$-.044$,.136] ($n = 231$) and .027 [$-.095$,.155] ($n = 74$); Phi-minus-Bio is .062 [$-.064$,.191] ($n = 65$). Each interval includes zero. Answer-bearing support is consistently helpful, whereas bridge-only prompting is model-dependent.

Path-specificity materially changes who is evaluated. At Qwen2.5-7B the bridge-exact and edge-exact gates retain 474 and 409 of the 970 licensed items; at Qwen3-8B they retain 491 and 395 of 885. Yet the support gap persists (Table 4): edge-exact estimates are .330 [.265,.397] and .228 [.167,.289]. The reduction for Qwen3-8B shows that set-valued admission inflated the licensed estimate, while the nonzero exact-edge result shows it did not create the entire phenomenon.

The gate-reliability curve in Table 5 tightens the free criterion while reusing unchanged Stage-2 predictions. Qwen2.5-7B shows no monotonic attenuation: gaps are .389/.380/.424/.369/.393 as the criterion moves from 1/5 to 5/5 per hop, although n falls from 970 to 28. Qwen3-8B declines from .341 to .174; its 5/5 interval remains above zero [.070,.283], but only 46 items remain. Thus full-fact sensitivity survives stable component retrieval at both endpoints, while its magnitude is not gate-invariant and the strictest estimands describe small selected populations.

5.3 Format and chain-of-thought sensitivity

For Qwen2.5-1.5B, 3B, and 7B, FULL-FACT−CoT is .322, .324, and .318, and FULL-FACT succeeds while both LATENT and CoT fail on 109/453, 197/829, and 227/970 items. The corrected CoT uses one 512-token output contract; Qwen3 native thinking is disabled, so this is an instruction effect rather than evidence of faithful reasoning (Wei et al., 2022; Kojima et al., 2022). Exact-term composed responses show the same support effect but larger gaps (.795 at Qwen2.5-7B and .727 at Qwen3-8B; Appendix Table 18), making output format part of the estimand.

5.4 MATCHED-SET COMPARISON ACROSS QWEN GENERATIONS

The conditional populations differ: Qwen2.5-7B admits 970 items and Qwen3-8B 885. On their 341 identical items, $\Delta_{\text{full}} = .349$ [.299,.402] and .343 [.293,.393]; the paired difference is .006 with bridge-cluster interval [-.055,.075] (Appendix Tables 15–16). A post-hoc $\pm.05$ equivalence margin is not met, while $\pm.10$ is; the data therefore do not resolve differences below roughly .08.

At the primary FULL-FACT threshold .80, five of seven checkpoints pass and only the 7B↔8B tier is comparable; .75 and .85 yield two and one tiers. Qwen3-1.7B is the boundary case (.775), but .80 remains primary.

5.5 GATE STABILITY, SPECIFICATION MIX, AND POPULARITY

The any-of-five free criterion is consequential. At Qwen2.5-7B, 6,455 items pass MCQ-only $4/5$ gates, 970 also pass at least one free paraphrase, 224 pass at least three, and 28 pass all five; the corresponding Qwen3-8B counts are 5,603, 885, 229, and 46. Table 5 shows that changing the threshold changes both population and effect size. A new independent probe bank uses five wordings per format and hop that never participate in admission. It retains 490/970 Qwen (50.5%), 346/589 Phi (58.7%), and 111/273 BioMistral items (40.7%). For Qwen, held-out MCQ accuracy remains high (.873/.937 for hops A/B), while free accuracy is .455/.459; bridge-exact and edge-exact retention fall to 238/970 and 207/970. The gate is therefore substantially prompt-conditioned.

The endpoint contrast is directionally more stable than admission. Five composed-question wordings authored after the original template was frozen yield positive FULL-FACT−LATENT gaps at every model–wording pair, with all 15 bridge-cluster intervals above zero. The gap ranges are .368–.546 for Qwen, .107–.324 for Phi, and .220–.363 for BioMistral (Appendix Table 13). The direction is robust in this bank, but the range confirms that its magnitude is prompt-dependent.

The gated sets are compositionally imbalanced. At Qwen2.5-7B, phenotype-class items account for 692/970 (71%), while disease-class has $n = 9$. Appendix Table 17 reports FULL-FACT−LATENT per specification with bridge-cluster intervals; disease-class remains underpowered. The support gap ranges from .230 for inheritance to .769 for gene–disease–phenotype at 7B, so a pooled score is not a relation-invariant capability measure.

Bridge degree changes absolute accuracy but does not eliminate the licensed support gap. For example, Qwen2.5-7B low/middle/high LATENT accuracies are .622/.588/.578, versus FULL-FACT .992/.994/.962; Qwen3-8B gives .679/.595/.684 versus 1.000/.980/.981. These are descriptive strata, not a causal popularity analysis.

6 DISCUSSION

Interpreting the full-fact gap. The licensed gate removes items with no demonstrated access to either component relation, and the exact-path variants remove cases supported by a different legal path. Across these populations, stating the chain improves both MCQ and free-response accuracy. At Qwen2.5-7B, the answer-bearing second fact accounts for 87.0% of the full-fact improvement, while bridge-only support is much smaller and forced chain elicitation leaves accuracy unchanged. Misleading facts redirect answers, and a FULL-FACT-length irrelevant context does not reproduce the gain. The experiment identifies an information-support effect; it does not specify the internal computation that produced the closed-book error.

Why model-specific conditioning matters. Conditioning improves interpretability but changes the population. A stronger model admits more and different items; the all-model intersection is empty. We therefore use S_m for within-model estimates and named intersections for cross-model comparisons. The same issue arises whenever an evaluation selects examples using the model being evaluated.

Benchmark design as an audit target. Sets prevent false negatives when several tails are valid, but a relation-level gate can cease to validate the sampled path. Following work on matched, versioned benchmark controls (Zhang et al., 2024; 2025), we expose licensed and exact-path definitions, request accounting, prompt stability, and cluster structure as parts of the measurement.

7 LIMITATIONS AND BROADER IMPACT

FULL-FACT contains the answer-bearing second-hop fact, so the primary contrast measures support sensitivity rather than composition alone. The licensed gate is prompt-sensitive and can admit a different legal path; the exact-path gates and composed-wording bank were added after the primary templates were frozen. A deterministic balanced sample, rather than the entire 18,856-item bank, supplies the full-population comparison, and no biomedical expert audit is available.

Model coverage is also limited. The seven-point sweep spans two Qwen generations, and the lineage check adds one Phi and one BioMistral endpoint; BioMistral admits 273 items, two of its specifications are underpowered, and no checkpoint exceeds 8B. Pairwise intersections contain 65–231 items and the three-way intersection contains 38. Within the benchmark, phenotype-class dominates the conditional sets and disease-class remains sparse. HPO and Mondo encode curated relations rather than clinical reasoning under uncertainty, and their public facts may have appeared in pretraining data. Greedy decoding evaluates one prompting regime; CoT measures the effect of an instruction with Qwen3 native thinking disabled.

Activation patching, attention knockout, and bridge readout each failed a design or positive-control audit. We make no mechanistic claim; Appendix F records the failure modes. The benchmark contains no patient data or treatment advice. Its main foreseeable misuse is treating conditional benchmark accuracy as evidence of clinical competence; the release counters this with provenance, per-specification results, and explicit control labels.

8 CONCLUSION

MEDCOMPBENCH shows that a cleaner denominator does not by itself identify a capability. Repeated component probes define a transparent conditional population, but set-valued answers can validate another path and independent wording changes who is admitted. The information controls trace most of the full-fact gain to the answer-bearing second fact; bridge-only support is smaller, and forced chain elicitation rarely recovers the sampled path. Passing component probes therefore does not make a multi-hop answer independent of explicit factual support. Conditional evaluations should report the gate, path criterion, prompt stability, dependency structure, and intervention contrast that define their claim.

REPRODUCIBILITY STATEMENT

Sections 3–4 and Appendix A–E document the bank construction, prompts, decoding, gate, grading, statistics, and compute environment. The submission package contains frozen aggregate records, table and figure exporters, and tests for deterministic construction, synonym and set-valued grading, branch-local cache keys, control exclusion, and strict JSON. The release checklist separately tracks anonymous hosting of the complete bank and graded records.

ETHICS STATEMENT

The benchmark is generated from public biomedical ontologies and contains no patient-level or personally identifiable data. It evaluates prompt-conditioned multi-hop behaviour rather than clinical decision making and must not be used to certify clinical competence or safety. Section 7 and Appendix D describe upstream licensing, ontology-coverage biases, intended use, and misuse mitigations.

AI USE STATEMENT

Generative AI tools assisted with experimental-design auditing, software implementation and debugging, dataset formatting and validation, aggregation and interpretation of results, literature discovery, and drafting and editing the manuscript. All benchmark definitions, model outputs, numerical results, statistical analyses, citations, and AI-assisted code were checked against frozen artifacts,

automated tests, and manual review. The authors take responsibility for the final content, including text, claims, and artifacts produced with AI assistance.

REFERENCES

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Pieter Gijsbers, Joan Giner-Miguelez, et al. Croissant: A metadata format for ML-ready datasets. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0590.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. To trust or not to trust? enhancing large language models' situated faithfulness to external contexts. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/186a213d720568b31f9b59c085a23e5a-Abstract-Conference.html.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421.

Apoorv Khandelwal and Ellie Pavlick. How do language models compose functions? *arXiv preprint arXiv:2510.01685*, 2025. doi: 10.48550/arXiv.2510.01685. URL https://arxiv.org/abs/2510.01685.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, et al. The human phenotype ontology in 2021. *Nucleic Acids Research*, 49 (D1):D1207–D1217, 2021. doi: 10.1093/nar/gkaa1043.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3381.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5848–5864, 2024. doi: 10.18653/v1/2024.findings-acl.348.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4249–4257, 2019. doi: 10.18653/v1/P19-1416.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. FaithEval: Can your language model stay faithful to context, even if the moon is made of marshmallows? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/48404cd9ce03946c6b7177691f3267a1-Abstract-Conference.html.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260, 2022.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP 2019*, pp. 2463–2473, 2019. doi: 10.18653/v1/D19-1250.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023. doi: 10.18653/v1/2023.findings-emnlp.378.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826, 2022. doi: 10.1145/3531146.3533231.

Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Nicole A. Vasilevsky, Nicolas Matentzoglu, Sabrina Toro, Joseph E. Flack, Harshad Hegde, Deepak R. Unni, et al. Mondo: Unifying diseases for the world, by the world. *medRxiv*, 2022. doi: 10.1101/2022.04.13.22273750.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025a. doi: 10.48550/arXiv.2505.09388.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3971–3992, 2025b. doi: 10.18653/v1/2025.findings-acl.205.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1485.

Liangliang Zhang, Zhuorui Jiang, Hongliang Chi, Haoyang Chen, Mohammed ElKoumy, Fali Wang, Qiong Wu, Zhengyi Zhou, Shirui Pan, Suhang Wang, and Yao Ma. Diagnosing and addressing pitfalls in KG-RAG datasets: Toward more reliable benchmarking. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2274.

A PROMPT AND GRADING SPECIFICATION

Frozen branch contracts. The prerequisite relations each have five paraphrases, indexed 0–4 and never reordered. Every paraphrase is rendered once as multiple choice and once as free response, giving ten requests per hop and twenty per real item. Multiple-choice probes end with "Answer with the single letter of the best option and nothing else." Free probes end with "Answer with the exact term only, no explanation, no punctuation." The LATENT branch adds "Give the final answer directly. Do not show any reasoning." The version-2 COT contract is "Think step by step. On the last line, write exactly `Final answer: <single option letter>`" for MCQ and the analogous exact-term instruction for free response. The FULL-FACT branch (artifact key `ORACLE`) prepends:

> Use only the following two established facts.
> Fact 1: [hop-A statement]
> Fact 2: [hop-B statement]

and then renders the same composed question and options.

Example. For a gene–disease–inheritance item, PROBE-A asks which disorder is caused by pathogenic variants in a gene; PROBE-B asks the inheritance pattern of the resulting disease. The composed question asks the inheritance pattern associated with the gene. LATENT omits the disease name, whereas FULL-FACT states both the gene–disease and disease–inheritance facts. The second statement explicitly contains the correct inheritance answer; this is why the branch is a full-fact ceiling rather than a bridge-only intervention. The answer options are identical across branches.

Synonyms and sets. Free responses are normalized for case, whitespace, and terminal punctuation, then checked against the preferred label and ontology synonyms. For set-valued chains, correctness is membership in the full legal-tail set, not equality to one sampled representative. This makes the implemented gate relation-valid but not necessarily path-specific. Of 11,643 disease–phenotype–phenotype-class items, 10,978 have multiple legal hop-A tails; 315 generated hop-A MCQs contain more than one ontology-valid option. The bridge-exact audit requires the current item bridge on hop A, while edge-exact also requires the sampled hop-B tail. Multiple-choice distractors are constructed from the item's local ontology neighbourhood with an item-ID-derived seed.

B IDENTIFICATION-SUITE SPECIFICATION AND RESULTS

The identification suite leaves the bank, gate, grader, frozen templates, and historical results unchanged. For every selected item, BRIDGE-ONLY supplies "The intermediate entity is [bridge]"; FACT-1 supplies the rendered hop-A assertion; FACT-2 supplies the rendered hop-B assertion; and FULL-FACT supplies both. FORCED-CHAIN uses the latent question and requires one `Bridge: <entity>` line followed by `Final answer: <letter>`. The wrong-bridge branch substitutes a same-relation donor bridge and its continuation; donors that make an option ambiguous are excluded. The wrong-second-hop branch replaces only the tail. LENGTH-CONTROL uses an unrelated two-fact context whose complete prompt exactly matches the FULL-FACT token count. Short branches retain the 32-token budget, FORCED-CHAIN uses 64, and COT uses 512. The cache key includes the model, prompt, and branch-local decoding parameters.

The composed-wording validation is post hoc but fixed before generation. For each frozen conditional item it renders five additional chain-specification-specific MCQ wordings in both LATENT and FULL-FACT form. The four option identities, order, answer set, decoder, and scoring are unchanged, giving ten new requests per item. Every wording is reported; no wording is selected after observing results. Intervals resample bridge clusters and exact McNemar values are Holm-adjusted across the five wordings within each checkpoint.

Model	Population	n	LATENT	Bridge	Fact 1	Fact 2	Forced	FULL-FACT	Wrong bridge	Wrong hop 2	Length ctrl.
BioMistral-7B	conditional	273	.707	.546	.637	.894	.304	.908	.040	.128	.370
BioMistral-7B	balanced bank	1600	.454	.236	.232	.796	.236	.836	.020	.032	.325
Qwen2.5-7B	conditional	970	.594	.691	.684	.932	.586	.982	.016	.352	.546
Qwen2.5-7B	balanced bank	1600	.502	.397	.369	.905	.491	.976	.004	.096	.421
Phi-3.5-mini	conditional	589	.643	.428	.380	.851	.557	.961	.015	.212	.587
Phi-3.5-mini	balanced bank	1600	.447	.228	.166	.726	.419	.805	.001	.032	.434

Table 6: Identification controls. Conditional rows use each model's licensed gate; balanced-bank rows use the frozen 400-per-specification sample. Wrong-bridge entries use the valid-donor subset; other branches share the stated denominator. FULL-FACT is the frozen artifact branch named ORACLE.

Model	Population	n	Bridge exact	Licensed bridge	Final correct
BioMistral-7B	conditional	273	111 (40.7%)	114 (41.8%)	83 (30.4%)
BioMistral-7B	balanced bank	1600	439 (27.4%)	446 (27.9%)	377 (23.6%)
Qwen2.5-7B	conditional	970	126 (13.0%)	202 (20.8%)	568 (58.6%)
Qwen2.5-7B	balanced bank	1600	117 (7.3%)	132 (8.2%)	786 (49.1%)
Phi-3.5-mini	conditional	589	113 (19.2%)	128 (21.7%)	328 (55.7%)
Phi-3.5-mini	balanced bank	1600	229 (14.3%)	232 (14.5%)	670 (41.9%)

Table 7: Forced-chain decomposition. The model must emit an explicit `Bridge:` line followed by one `Final answer:` letter. Exact uses the sampled path entity; licensed accepts any ontology-valid first-hop tail.

Model	n	A-MCQ	A-free	B-MCQ	B-free	Retained
BioMistral-7B	273	.903	.473	.758	.406	111 (40.7%)
Qwen2.5-7B	970	.873	.455	.937	.459	490 (50.5%)
Phi-3.5-mini	589	.959	.386	.893	.402	346 (58.7%)

Table 8: Independent held-out prerequisite probes. Each hop uses five newly worded MCQ and five newly worded free-response prompts; none participates in the original gate. Retained requires the unchanged $4/5$ MCQ plus any-of-five licensed-free rule on both hops.

Model	Spec.	n	Bridge−LATENT	Fact1−LATENT	Fact2−LATENT	FULL-FACT−LATENT	Length−LATENT
BioMistral-7B	DP–PC	91	−.407 [−.585,−.284]	−.275 [−.389,−.176]	.154 [.016,.245]	.132 [−.029,.239]	−.033 [−.144,.065]
BioMistral-7B	GD–DC†	1	.000 [.000,.000]	.000 [.000,.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]	.000 [.000,.000]
BioMistral-7B	GD–INH	155	−.032 [−.090,.032]	.058 [.000,.116]	.110 [.045,.174]	.142 [.084,.200]	−.568 [−.645,−.490]
BioMistral-7B	GD–PH†	26	−.077 [−.296,.120]	−.115 [−.320,.077]	.731 [.556,.889]	.769 [.600,.923]	−.038 [−.269,.192]
Qwen2.5-7B	DP–PC	692	.068 [.018,.118]	.081 [.028,.135]	.328 [.260,.400]	.397 [.325,.471]	−.004 [−.053,.049]
Qwen2.5-7B	GD–DC†	9	.000 [−.333,.333]	−.111 [−.444,.222]	.444 [.111,.778]	.556 [.222,.889]	.111 [.000,.333]
Qwen2.5-7B	GD–INH	204	.181 [.125,.240]	.206 [.146,.270]	.230 [.173,.290]	.230 [.173,.290]	−.201 [−.290,−.113]
Qwen2.5-7B	GD–PH	65	.154 [.031,.277]	−.154 [−.262,−.047]	.769 [.667,.869]	.769 [.672,.869]	−.046 [−.151,.065]
Phi-3.5-mini	DP–PC	451	−.290 [−.376,−.202]	−.319 [−.410,−.215]	.195 [.087,.293]	.333 [.267,.404]	−.011 [−.071,.058]
Phi-3.5-mini	GD–DC†	1	.000 [.000,.000]	.000 [.000,.000]	.000 [.000,.000]	.000 [.000,.000]	.000 [.000,.000]
Phi-3.5-mini	GD–INH	114	.035 [−.035,.109]	−.070 [−.149,.009]	.158 [.090,.231]	.167 [.101,.239]	−.237 [−.336,−.142]
Phi-3.5-mini	GD–PH†	23	.000 [−.174,.174]	−.130 [−.261,.000]	.696 [.478,.870]	.783 [.609,.957]	−.043 [−.174,.087]

Table 9: Per-specification identification contrasts on each model's conditional set. DP–PC: disease→phenotype→phenotype class; GD–DC: gene→disease→disease class; GD–INH: gene→disease→inheritance; GD–PH: gene→disease→phenotype. Entries are paired differences with bridge-cluster 95% stability intervals. †Underpowered ($n < 50$).

Table 10: Identification controls on the 231 items admitted by both Qwen2.5-7B and Phi-3.5-mini. Entries are paired accuracy differences from LATENT with bridge-cluster 95% stability intervals. The last column is the paired difference of the two model contrasts; wrong-bridge uses the 99-item common valid-donor subset.

Intervention − LATENT	Qwen2.5-7B	Phi-3.5-mini	Qwen − Phi
FULL-FACT	.333 [.249,.423]	.294 [.214,.383]	.039 [−.044,.136]
Bridge only	.117 [.050,.189]	−.234 [−.349,−.110]	.351 [.226,.458]
First fact only	.130 [.058,.203]	−.260 [−.367,−.146]	.390 [.278,.495]
Second fact only	.294 [.220,.378]	.165 [.037,.283]	.130 [.021,.251]
Forced chain	−.030 [−.084,.022]	−.091 [−.148,−.040]	.061 [−.019,.143]
Wrong bridge	−.677 [−.784,−.566]	−.687 [−.816,−.556]	.010 [−.096,.129]
Wrong second hop	−.190 [−.302,−.089]	−.455 [−.558,−.356]	.264 [.155,.370]
Length-matched irrelevant	−.009 [−.074,.054]	−.091 [−.174,−.004]	.082 [−.004,.172]

Left checkpoint	Right checkpoint	n	Left gap	Right gap	Difference
Qwen2.5-7B	Phi-3.5-mini	231	.333 [.249, .423]	.294 [.214, .383]	.039 [−.044, .136]
Qwen2.5-7B	BioMistral-7B	74	.257 [.160, .362]	.230 [.135, .329]	.027 [−.095, .155]
Phi-3.5-mini	BioMistral-7B	65	.231 [.129, .344]	.169 [.049, .281]	.062 [−.063, .190]

Table 11: Pairwise matched-set FULL-FACT−LATENT comparisons. Each row contains items admitted by both checkpoints. Difference is left minus right on identical items; intervals resample bridge clusters.

Model	n	FULL-FACT	Bridge	Fact 2	Wrong bridge	Length ctrl.
Qwen2.5-7B	38	.263 [.128, .421]	.237 [.108, .385]	.263 [.128, .417]	−.840 [−.963, −.680]	−.079 [−.243, .086]
Phi-3.5-mini	38	.184 [.073, .316]	−.053 [−.194, .095]	.158 [.027, .297]	−.760 [−.920, −.565]	−.237 [−.389, −.083]
BioMistral-7B	38	.263 [.135, .400]	−.105 [−.244, .029]	.211 [.098, .333]	−.520 [−.720, −.333]	−.184 [−.333, −.049]

Table 12: Information-control contrasts on the intersection admitted by all three independently developed checkpoints ($n = 38$). Entries are paired accuracy differences from LATENT with bridge-cluster 95% stability intervals. Wrong-bridge and exactly length-matched controls use their available common subsets, reported in the machine-readable record.

Model	Wording	n	LATENT	FULL-FACT	Gap	95% stability interval
Qwen2.5-7B	H1	970	.620	.988	.368	[.310, .426]
	H2	970	.552	.980	.429	[.360, .496]
	H3	970	.424	.970	.546	[.480, .609]
	H4	970	.559	.996	.437	[.376, .496]
	H5	970	.542	.960	.418	[.352, .480]
Phi-3.5-mini	H1	589	.598	.822	.224	[.156, .299]
	H2	589	.538	.839	.301	[.236, .365]
	H3	589	.576	.900	.324	[.270, .384]
	H4	589	.603	.786	.183	[.097, .270]
	H5	589	.474	.581	.107	[.028, .189]
BioMistral-7B	H1	273	.634	.853	.220	[.150, .285]
	H2	273	.586	.872	.286	[.224, .351]
	H3	273	.421	.784	.363	[.280, .438]
	H4	273	.505	.817	.311	[.240, .382]
	H5	273	.608	.835	.227	[.145, .306]

Table 13: Independent composed-question wording validation on each model's frozen conditional set. H1–H5 were authored after the primary templates were frozen. Option identities and order are unchanged. Intervals resample bridge clusters; these post-hoc wordings do not alter admission or the primary estimand.

The balanced bank contains a deterministic sample of 400 items per chain specification. Its denominator remains separate from the model-conditional sets. The Qwen2.5-7B evaluation union has 2,507 unique items because the 1,600 balanced items overlap its 970-item conditional set. All 2,910 frozen conditional LATENT/COT/ORACLE requests were cache hits. Phi-3.5-mini and BioMistral-7B admit 589 and 273 items. Their pairwise intersections with Qwen contain 231 and 74 items; Phi and BioMistral share 65, and all three models share 38. Tables 10–12 report model-specific and matched-item estimates separately. The expert-review packet contains 75 deterministic items per specification; no human ratings are available in this revision.

C SUPPLEMENTARY BEHAVIOURAL RESULTS

C.1 POOLED AND MACRO ESTIMATES

Table 14: Pooled and equal-spec macro CCA. Pooled scores weight items; macro scores weight equally the specifications with $n_{\text{spec}} \geq 50$. The number of included specifications is shown in parentheses. Sparse 7B/8B specifications remain visible in Table 17 but do not enter the macro.

	Pooled			Macro		
Model	LATENT	CoT	FULL-FACT	LATENT	CoT	FULL-FACT
Qwen2.5-0.5B	.403	.434	.453	.389 (1)	.420 (1)	.382 (1)
Qwen2.5-1.5B	.647	.651	.974	.622 (2)	.714 (2)	.986 (2)
Qwen2.5-3B	.574	.642	.966	.588 (2)	.751 (2)	.977 (2)
Qwen2.5-7B	.594	.665	.982	.527 (3)	.579 (3)	.993 (3)
Qwen3-0.6B	.589	.536	.848	.603 (1)	.441 (1)	.824 (1)
Qwen3-1.7B	.505	.363	.775	.522 (2)	.501 (2)	.851 (2)
Qwen3-8B	.646	.678	.988	.570 (3)	.615 (3)	.993 (3)

Table 15: Matched-tier branch contrasts on the identical licensed subset S_2 ($n = 341$). Intervals are paired item bootstrap; exact McNemar–Holm values are retained as legacy diagnostics but do not account for entity clustering.

	Qwen2.5-7B	Qwen3-8B
FULL-FACT−LATENT	.349 [.299, .402]	.343 [.293, .393]
Holm p	6.0×10^{-36}	1.2×10^{-35}
FULL-FACT−CoT	.243 [.196, .290]	.264 [.217, .311]
Holm p	4.4×10^{-24}	3.2×10^{-27}

Table 16: Direct paired comparison of the two full-fact support gaps on the identical S_2 items. The estimand is $(\Delta^{2.5}_{\text{full}}) - (\Delta^{3}_{\text{full}})$. The post-hoc equivalence checks are sensitivity analyses, not preregistered tests.

Quantity	Estimate
Paired difference ($n = 341$)	.006
Item-bootstrap 95% stability interval	[-.050, .065]
Head-cluster 95% stability interval	[-.050, .062]
Bridge-cluster 95% stability interval	[-.055, .075]
Post-hoc equivalence, margin $\pm.05$	not established
Post-hoc equivalence, margin $\pm.10$	established

The macro columns give equal weight to specifications with $n \geq 50$ and report how many specifications enter each value. With only one to three eligible specifications, these are descriptive summaries; uncertainty for the primary paired gaps is reported through bridge-cluster stability intervals.

C.2 PER-SPECIFICATION STRATA

Table 17: Per-specification FULL-FACT support gaps for the 7B/8B endpoints under the licensed gate. Intervals use bridge-entity cluster bootstrap. Rows with $n < 50$ are retained but underpowered.

Model	Specification	n	LATENT	FULL-FACT	Gap [95% stability interval]	Status
Qwen2.5-7B	disease→phenotype→class	692	.581	.978	.397 [.324, .471]	included
	gene→disease→class	9	.222	.778	.556 [.222, .889]	underpowered
	gene→disease→inheritance	204	.770	1.000	.230 [.175, .290]	included
	gene→disease→phenotype	65	.231	1.000	.769 [.671, .869]	included
Qwen3-8B	disease→phenotype→class	556	.637	.980	.344 [.270, .417]	included
	gene→disease→class	5	.400	1.000	.600 [.200, 1.000]	underpowered
	gene→disease→inheritance	225	.871	1.000	.129 [.088, .174]	included
	gene→disease→phenotype	99	.202	1.000	.798 [.714, .876]	included

The gene–disease–phenotype endpoint remains difficult even when it reaches the $n \geq 50$ floor: at 7B/8B, LATENT is .231/.202 while FULL-FACT is 1.000 in both frozen metrics. Disease-class is retained but never used for a trend claim because every model has $n < 50$ (Qwen2.5-7B has $n = 9$).

C.3 COMPOSED FREE RESPONSE

Table 18: Composed free-response accuracy on the licensed conditional sets. The final column reports FULL-FACT−LATENT with a bridge-cluster stability interval.

Model	n	LATENT	CoT	FULL-FACT gap [95% stability interval]
Qwen2.5-0.5B	159	.390	.258	.308 [.190, .412]
Qwen2.5-1.5B	453	.128	.137	.865 [.823, .902]
Qwen2.5-3B	829	.191	.201	.808 [.767, .843]
Qwen2.5-7B	970	.204	.229	.795 [.752, .831]
Qwen3-0.6B	112	.232	.357	.696 [.581, .807]
Qwen3-1.7B	311	.283	.212	.701 [.603, .783]
Qwen3-8B	885	.271	.330	.727 [.680, .767]

Each free-response condition uses one exact-term composed prompt. Its full-fact support gap is larger than the MCQ gap, so the two formats are reported separately.

C.4 COUNTERFACTUAL CEILING

Table 19: Counterfactual in-context binding accuracy. These rows form a separate control population and do not enter n_{cond}.

Model	CF accuracy [95% item-resampling interval]	n
Qwen2.5-0.5B	.579 [.536, .621]	530
Qwen2.5-1.5B	.938 [.918, .957]	600
Qwen2.5-3B	.858 [.831, .884]	688
Qwen2.5-7B	.943 [.927, .960]	723
Qwen3-0.6B	.677 [.516, .839]	31
Qwen3-1.7B	.988 [.963, 1.000]	80
Qwen3-8B	1.000 [1.000, 1.000]	217

C.5 POPULARITY BINS

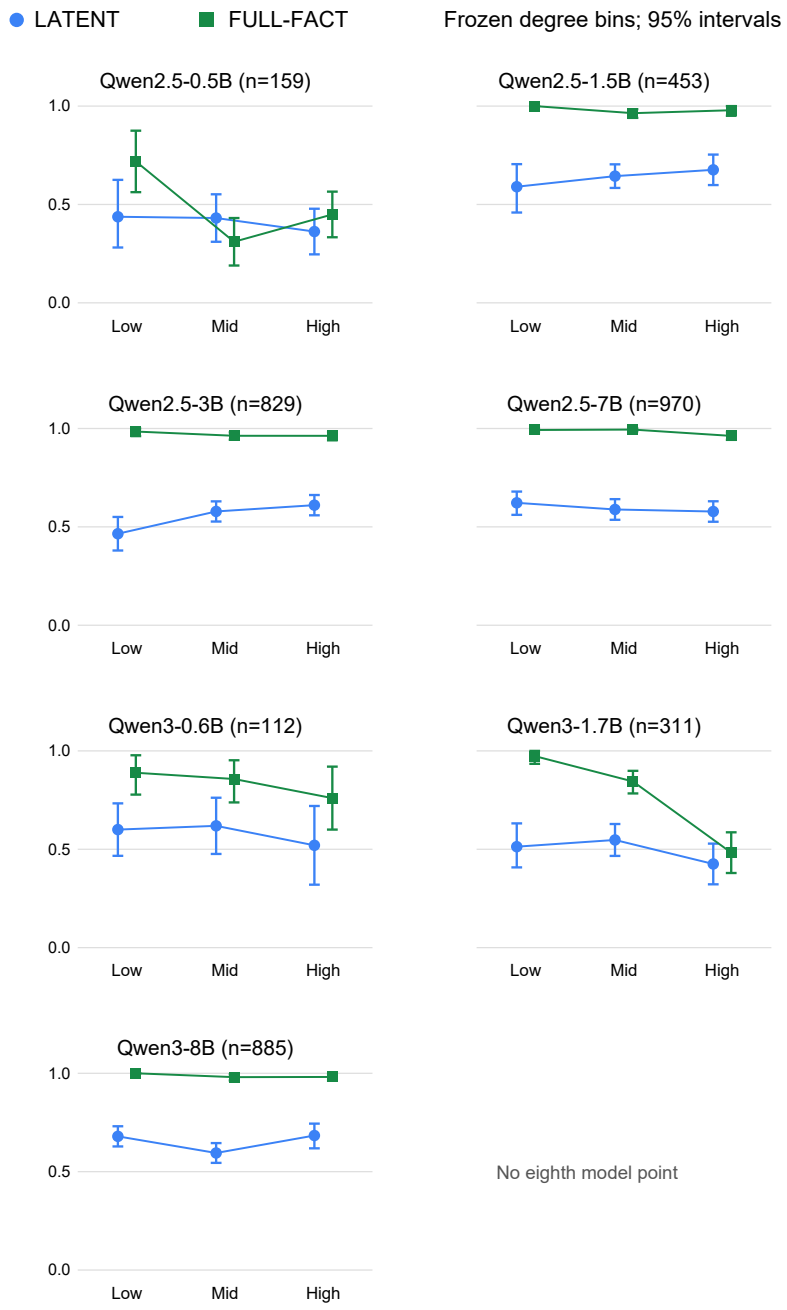


Figure 2: LATENT and FULL-FACT accuracy across frozen bridge-degree bins. Bins are defined on all real items before gating; error bars are the frozen item-bootstrap intervals.

C.6 FREE-GATE RELIABILITY ACROSS ALL CHECKPOINTS

Table 20: Full post-hoc free-gate curve. Cells show the multiple-choice full-fact support gap with the admitted n in parentheses. Bridge-cluster intervals for every cell are stored in the accompanying audit JSON; the two endpoint models are shown with intervals in Table 5.

Model	1/5	2/5	3/5	4/5	5/5
Qwen2.5-0.5B	.050 (159)	.000 (68)	.062 (16)	.000 (6)	.000 (1)
Qwen2.5-1.5B	.327 (453)	.281 (160)	.269 (78)	.357 (14)	.000 (1)
Qwen2.5-3B	.392 (829)	.378 (296)	.407 (108)	.412 (34)	.333 (3)
Qwen2.5-7B	.389 (970)	.380 (468)	.424 (224)	.369 (111)	.393 (28)
Qwen3-0.6B	.259 (112)	.294 (34)	.444 (18)	.500 (6)	– (0)
Qwen3-1.7B	.270 (311)	.208 (48)	.200 (10)	.333 (3)	– (0)
Qwen3-8B	.341 (885)	.324 (423)	.306 (229)	.282 (131)	.174 (46)

The strictest thresholds can leave fewer than five items, so their point estimates are descriptive rather than stable comparisons. The endpoint analysis in Table 5 reports every bridge-cluster interval and anchors the interpretation on the larger 7B/8B sets.

C.7 FULL-FACT-CONTROL THRESHOLD SENSITIVITY

Table 21: Sensitivity of the FULL-FACT control rule. The fixed main threshold remains 0.80; 0.75 and 0.85 are sensitivity analyses only. Parentheses give model-specific n_{cond}.

FULL-FACT threshold	Eligible model points (n_{cond})	Comparable tiers	Status
0.75	Qwen2.5-1.5B (453), 3B (829), 7B (970); Qwen3-0.6B (112), 1.7B (311), 8B (885)	2	sensitivity
0.80	Qwen2.5-1.5B (453), 3B (829), 7B (970); Qwen3-0.6B (112), 8B (885)	1	primary
0.85	Qwen2.5-1.5B (453), 3B (829), 7B (970); Qwen3-8B (885)	1	sensitivity

The fixed primary threshold is .80. Lowering it to .75 admits Qwen3-1.7B, whose FULL-FACT accuracy is .775; the threshold table therefore reports this checkpoint as a boundary case. The .75 and .85 rows are sensitivity analyses.

D DATA DOCUMENTATION

Motivation and intended use. MEDCOMPBENCH studies multi-hop behaviour after a model passes component-relation probes. The full-fact contrast measures support sensitivity rather than isolated composition. The resource is intended for language-model evaluation and benchmark-design research, not diagnosis, treatment, or clinical certification. The release follows dataset-card and datasheet principles (Gebru et al., 2021; Pushkarna et al., 2022) and includes Croissant metadata (Akhtar et al., 2024).

Composition and instances. The release contains 18,856 real two-hop chains and 4,711 counterfactual twins. Each real record stores stable entity identifiers and labels, the two edges, the legal answer set, four options, semantic-distance and popularity metadata, construction anchors, ontology provenance, and counterfactual linkage where applicable. A synonym sidecar maps valid ontology aliases. No patient-level or personal data are present.

Collection and preprocessing. The bank is generated programmatically from HPO and Mondo releases dated 2025-05-06. Construction uses seed 20260824, fixed quotas, relation-specific validity rules, and deterministic local-neighbourhood distractors. The release manifest records source hashes. Expanding a quota was regression-tested to preserve the complete earlier item prefix field-for-field.

Licensing and attribution. HPO is used under its custom reuse terms, which require consortium attribution, version display, and preservation of the ontology content and logical relations. Mondo

is CC BY 4.0. Qwen2.5-0.5B, 1.5B, and 7B and all evaluated Qwen3 checkpoints are Apache 2.0; Qwen2.5-3B is under the Qwen Research License. The derived benchmark is distributed with the upstream notices and an HPO-compatible data-use notice rather than being represented as unrestricted raw ontology data.

Maintenance and versioning. The version string pins both ontology dates. New ontology releases must produce a new bank version; items and result folders are immutable. Deprecated or corrected upstream relations are not silently edited in an existing version. Issues should identify the item ID, source edge, and ontology release.

Known limitations. The bank reflects ontology curation coverage and inherits its historical and geographic biases. Chain specifications are imbalanced, public ontology facts may occur in model pretraining data, and label aliases are more complete for some entities than others. Set-valued answers reduce false negatives but do not establish that every medically plausible answer absent from the pinned ontologies is wrong.

E REPRODUCIBILITY AND COMPUTE

The reviewer package contains the bank manifest, configurations, prompt templates, frozen aggregate records, aggregation code, and table/figure exporters. The workflow comprises deterministic bank construction, Stage-1 probe evaluation, Stage-2 conditional evaluation, and metric export. Unit and acceptance tests cover quota-prefix stability, synonym and set-valued grading, branch-local cache keys, empty macro buckets, ORACLE control exclusion, strict JSON, and prompt-version separation. The complete bank and per-model graded records remain a separate anonymous-hosting requirement.

All model inference used one cloud GPU: an NVIDIA GeForce RTX 4090 with 24 GB memory. The host ran Ubuntu 22.04 with Python 3.12, PyTorch 2.8.0, and CUDA 12.8; it provided 16 vCPUs from an Intel Xeon Platinum 8352V at 2.10 GHz, 120 GB system memory, a 30 GB system disk, and a 50 GB SSD data disk. All weights used bfloat16 and no quantization. Peak recorded allocation was 23,718 MiB for the Qwen2.5-7B behavioural Stage 2. Its Stage 1 took 3,117 s for 300,125 uncached backend generations (96.51/s, including model load, grading, and writes), and Stage 2 took 313 s. Qwen3 total measured wall times were 1,574.8 s (0.6B), 2,541.2 s (1.7B), and 4,314.7 s (8B), with peak observed memory 22,728, 21,914, and 22,442 MiB. Each Qwen3 Stage 1 contains exactly 377,120 requests: $18{,}856 \times 2 \times 5 \times 2$. Stage-2 counts are 734, 2,026, and 5,744; the 8B count decomposes as $885 \times 3 \times 2 + 217 \times 2$. Qwen2.5 Stage 1 evaluates 16,902 extension items (338,040 rows) and merges 39,080 frozen rows for the old 1,954 items, again totaling 377,120 effective probe rows. Historical Qwen2.5-0.5B/1.5B/3B Stage-1 wall times were not preserved and are reported as missing rather than estimated.

The post-audit Qwen2.5-7B identification run issued 43,082 requests, of which 4,272 were cache hits; all 2,910 required conditional LATENT/COT/FULL-FACT requests were present. Its 38,810 backend generations took 559.1 s. Phi-3.5-mini Stage 1 issued 377,120 uncached probe requests and admitted 589 items; the Stage-1 wall time and peak allocation were not persisted. Its subsequent identification run issued 32,131 uncached requests in 526.0 s. BioMistral-7B Stage 1 also issued 377,120 requests: 131,371 survived the interrupted allocation and 245,749 were generated after resumption. The split run's end-to-end wall time and peak allocation were not persisted. It admitted 273 items; the 23,000-request identification suite then ran uncached in 592.8 s. The independent composed-wording runs issued 9,700/5,890/2,730 requests for Qwen/Phi/BioMistral. Backend generation took 104.9/55.0/47.2 s; end-to-end times were 139.5/553.8/85.1 s, where Phi includes 446.0 s to restore a missing local snapshot from the official checkpoint.

F MECHANISTIC AUDIT RECORD

Three mechanistic routes were tested on Qwen2.5-7B; none passed the required design and positive-control checks.

1. An attention-knockout implementation was initially coupled to an attention backend, and its nonce control shifted token positions by 11. Corrected continuous-margin tests still lacked a validated option-span control, so the route is not interpretable.
2. Two activation-patching studies used invalid interventions: the first corrupted the answer rather than the head and left the intended bridge unchanged; the second assigned a bridge anchor in a LATENT prompt that contained no bridge string. Their null effects were withdrawn. A separate single-hop subject-patching positive control passed ($n = 60$, peak layer 7, recovery .976 [.918, 1.027], Holm $p = 4.66 \times 10^{-10}$), validating only that patching path.
3. A raw-logit bridge readout failed its ORACLE positive control: the explicit bridge entered the top-100 at the prespecified shallow location for only 10/60 probes (16.7%). The main readout was therefore not run.

The resulting category is "pipeline not validated in this setting." The archived runs serve as a methodological record and as regression targets; they provide no bridge-specific mechanism result.