Supplementary Information

Candidate information structure reduces over-alignment in LLM-based concept mapping

This document accompanies the manuscript of the same title and provides (i) candidate-set construction details and dataset statistics, (ii) the full experimental prompts and decoding configuration, (iii) supplementary statistical results referenced in the main text, and (iv) supplementary figures. Table and figure numbering continues from the main manuscript (Tables S1–S8; Figures S1–S2).

S1. Candidate-Set Construction and Dataset Statistics

S1.1. Construction procedure.

Candidate sets C(t) for both domains were constructed by a unified procedure combining retrieval recall with expert verification. For each source term t, an initial pool of candidate concepts was drawn from the domain ontology (Domain A: the CMathAlign-derived mathematical concept schema; Domain B: the ICD-10 standard vocabulary), filtered for surface-form overlap and semantic plausibility, and then manually verified by domain experts so that every candidate set satisfied the following pre-registered constraints: (i) gold-blind construction—candidates were assembled without reference to the gold label; (ii) coverage = 100%—each C(t) contains at least one string that exact_match (including the synonym table) judges correct, so the Hard-Constraint ceiling is not artificially depressed; and (iii) accidental-hit rate < 50%—after permuting gold labels, the reconstructed coverage remains below 50%, confirming that C(t) does not trivially enclose everything. Candidate difficulty alignment across domains was assessed from the distribution of gold_rank_in_C and candidate-set size rather than from qualitative assertion alone (Table S1).

S1.2. Candidate-set statistics.

Table S1 summarizes candidate-set size |C| and the rank of the gold answer within the model's candidate ranking (gold_rank_in_C) for the 150 term instances in each domain. gold_rank_in_C is derived from the teacher-forcing log-probability ordering described in the main text (Section 3.3).

Table S1. Candidate-set size and gold-rank distribution for the two controlled-probe domains.

Statistic	Domain A (n = 150)	Domain B (n = 150)
Candidate-set size |C| — min / max	50 / 101	50 / 84
|C| — mean (SD)	72.1 (23.4)	64.3 (8.8)
|C| — median	53.5	64.0
gold_rank_in_C — min / max	1 / 97	1 / 77
gold_rank_in_C — mean (SD)	20.9 (23.6)	22.3 (19.8)
gold_rank_in_C — median (IQR)	11 (4–28)	14 (5–38)
gold_rank_in_C ≤ 5 — n (%)	49 (32.7%)	42 (28.0%)

gold_rank_in_C 6–10 — n (%)	21 (14.0%)	18 (12.0%)
gold_rank_in_C 11–20 — n (%)	21 (14.0%)	24 (16.0%)
gold_rank_in_C 21–50 — n (%)	42 (28.0%)	50 (33.3%)
gold_rank_in_C > 50 — n (%)	17 (11.3%)	16 (10.7%)
gold_share — mean / median	0.057 / 0.026	0.041 / 0.020
norm_mass — mean (SD) / median	−0.116 (2.45) / 0.535	−0.353 (2.29) / 0.126

Difficulty composition of the controlled subsets: Domain A comprised 150 pre-modern Chinese mathematical term instances stratified as L0/L1/L2/L3 = 46/41/33/30; Domain B comprised 150 Chinese clinical term instances mapped to ICD-10 with mapping types equivalence (n = 125) and superordinate generalization (n = 25). The over-alignment diagnostic core set comprised 828 terms (L0/L1/L2/L3 = 158/378/111/181).

S2. Experimental Prompts and Decoding Configuration

All four conditions (UC, HC, RO, SC) were administered through constrained decoding (vLLM with the xgrammar backend), in which the system and user prompts were held constant and only the guided-choice candidate list differed across conditions. Chinese prompts are reproduced verbatim with English glosses.

S2.1. Domain A (pre-modern Chinese mathematical terminology).

System prompt:

(English: “You are an expert in ancient Chinese mathematical literature.”)

User prompt template:

{term}
: {source}
: {definition}

:

(English: “Please provide the modern mathematical concept corresponding to the classical term ‘{term}’.
Source: {source}
Definition: {definition}

Please output only the mapping result, in the format:
Mapping result: ”)

S2.2. Domain B (Chinese clinical terminology → ICD-10).

System prompt:

(English: “You are an expert in clinical terminology standardization.”)

User prompt template:

{raw_term} ICD-10
: {context}

:

(English: “Please provide the ICD-10 standard diagnostic term corresponding to the clinical diagnosis term ‘{raw_term}’.
Context: {context}

Please output only the standard term, in the format:
Standard term: ”)

S2.3. Decoding configuration.

The four-condition probe and the candidate_mass computation used Qwen/Qwen3-32B (locally deployed, bfloat16) with vLLM decoding temperature = 0, top_p = 1.0, max_tokens = 32, and thinking disabled. Candidate randomization used seed 20260101; statistical resampling used seed 42 with 10,000 paired bootstrap iterations. Because decoding was deterministic (temperature = 0), the three repetitions per term produced identical outputs; the observed repetition consistency was 100% in both domains and all conditions, so majority-vote accuracy at the term level equals the single-output accuracy reported in the main text.

S3. Supplementary Statistical Results

S3.1. Per-condition accuracy and repetition consistency.

Table S2 reports the term-level majority-vote accuracy for each condition and domain. The three repetitions (rep0–rep2) were identical in every condition (100% consistency), reflecting deterministic decoding.

Table S2. Majority-vote accuracy by condition (term-level mean of three repetitions).

Condition	Domain A	Domain B
UC (Unconstrained)	0.033	0.040
HC (Hard Constraint)	0.273	0.467
RO (Reranked Order)	0.567	0.767
SC (Shuffled Constraint)	0.260	0.627

S3.2. Full pairwise comparisons and Holm-corrected p-values.

Table S3 reports all six pairwise condition contrasts (majority-vote accuracy, term level) with 95% bootstrap confidence intervals (10,000 paired resamples) and McNemar p-values on the binarized majority-vote outcomes. Holm correction was applied across the twelve comparisons (six contrasts × two domains). All significant comparisons remained significant after correction; the single non-significant contrast (SC−HC in Domain A) is reported unchanged.

Table S3. Pairwise condition contrasts with 95% bootstrap CIs and Holm-corrected McNemar p-values.

Contrast	A: Δ [95%	A: p	A: p (Holm)	B: Δ [95%	B: p	B: p (Holm)

	CI]	(McNemar)		CI]	(McNemar)	
RO − HC	+0.293 [0.220, 0.373]	<0.001	<0.001	+0.300 [0.227, 0.380]	<0.001	<0.001
HC − UC	+0.240 [0.167, 0.313]	<0.001	<0.001	+0.427 [0.347, 0.507]	<0.001	<0.001
RO − SC	+0.307 [0.233, 0.387]	<0.001	<0.001	+0.140 [0.073, 0.207]	<0.001	<0.001
SC − UC	+0.227 [0.153, 0.307]	<0.001	<0.001	+0.587 [0.507, 0.667]	<0.001	<0.001
RO − UC	+0.533 [0.447, 0.613]	<0.001	<0.001	+0.727 [0.653, 0.793]	<0.001	<0.001
SC − HC	−0.013 [−0.087, 0.060]	0.837	0.837	+0.160 [0.093, 0.227]	<0.001	<0.001

S3.3. Entity-linking retriever baselines.

Table S4 reports the full evaluation metrics (top-1 accuracy, mean reciprocal rank MRR, recall@5, and mean gold rank) for the three retrievers on the same candidate sets used in the four-condition probe.

Table S4. Entity-linking retriever baselines (full metrics).

Method	Domain	Top-1	MRR	R@5	Mean gold rank
bge-m3 (560M, bi-encoder)	A	0.373	0.468	0.560	13.4
bge-m3 (560M, bi-encoder)	B	0.573	0.672	0.767	5.9
bge-reranker-base (110M, cross-encoder)	A	0.367	0.469	0.560	14.2
bge-reranker-base (110M, cross-encoder)	B	0.540	0.651	0.780	5.8
bge-reranker-v2-m3 (560M, cross-encoder)	A	0.407	0.508	0.633	14.2
bge-reranker-v2-m3 (560M, cross-encoder)	B	0.580	0.673	0.793	5.1

S3.4. Cross-domain comparisons.

Table S5 reports the cross-domain comparison of record-level constraint gains. The HC−UC gain differed significantly between domains (Mann–Whitney U test, $p = 0.0013$; Cliff's $\delta = 0.18$), whereas the RO−HC gain did not ($p = 0.92$; Cliff's $\delta = 0.007$), consistent with the claim that the

directional RO>HC advantage is robust while the benefit of a single correct hint is domain-dependent.

Table S5. Cross-domain comparison of record-level gains.

Gain	Domain A mean (SD)	Domain B mean (SD)	Mann–Whitney p	Cliff's δ
HC − UC	0.240 (0.472)	0.427 (0.508)	0.0013	0.18
RO − HC	0.293 (0.478)	0.300 (0.483)	0.92	0.007

S3.5. candidate_mass descriptive statistics and correlation.

Table S6 reports the candidate_mass descriptive statistics (norm_mass = log(gold_share × |C|)) and its association with HC−UC record-level gains. Neither Pearson nor Spearman correlation was significant in either domain, delimiting the explanatory scope of candidate_mass (Figure 5 of the main text).

Table S6. candidate_mass descriptive statistics and correlation with HC−UC gain.

Statistic	Domain A	Domain B
norm_mass — mean (SD)	−0.116 (2.447)	−0.353 (2.289)
norm_mass — median	0.535	0.126
Pearson r (HC−UC gain)	−0.095	−0.082
Pearson r 95% bootstrap CI	[−0.251, 0.058]	[−0.234, 0.080]
Pearson p (two-sided)	0.247	0.323
Spearman ρ (HC−UC gain)	−0.124	−0.070

S3.6. Regression: candidate_mass × domain interaction (H4).

Table S7 reports the ordinary least-squares regression of HC−UC record-level gain on candidate_mass, domain, and their interaction (n = 300 pooled records). The interaction term was not significant (p = 0.994), indicating that the null mass → gain relationship did not differ across domains; the significant domain main effect (p = 0.0014) reflects the domain-dependent HC−UC baseline gain reported in Table S5.

Table S7. OLS regression of HC−UC gain on candidate_mass, domain, and their interaction.

Term	Estimate (SE)	t	p
Intercept (Domain A baseline)	+0.238 (0.040)	+5.92	<0.001
candidate_mass (Domain A slope)	−0.018 (0.016)	−1.12	0.262
domain = B	+0.182 (0.057)	+3.19	0.0014
candidate_mass × domain = B	+0.0002 (0.024)	+0.01	0.994

Model R^2 = 0.043; adjusted R^2 = 0.033; n = 300.

S3.7. Three-model diagnostics: per-level accuracy and confidence calibration.

Table S8 reports, for the three commercial models, the per-level (L0–L3) accuracy under both the expert annotation and the ontology annotation, together with mean self-reported confidence and the expected calibration error (ECE). Under the ontology annotation all L3 terms are labelled "no equivalent", so the L3 ontology accuracy equals the NONE-output (correct-rejection) rate. The expert annotation retains a closest-MSC reference for 161 of 181 L3 terms, so the expert L3 accuracy reflects best-match agreement rather than rejection; this explains the divergence between the two scoring schemes at L3.

Table S8. Per-level accuracy and confidence calibration for the three commercial models (core set, n = 828).

Model / scoring	L0 (158)	L1 (378)	L2 (111)	L3 (181)	Overall	Mean conf.	ECE
DeepSeek-Chat (expert)	51.3	29.1	27.0	53.6	38.4	89.0	0.506
DeepSeek-Chat (ontology)	51.3	29.1	27.0	5.5	27.9	89.0	0.611
GPT-4o-mini (expert)	30.4	25.9	25.2	22.7	26.0	85.6	0.596
GPT-4o-mini (ontology)	29.7	25.9	25.2	11.0	23.3	85.6	0.623
Claude Haiku 4.5 (expert)	27.2	21.4	25.2	24.3	23.7	77.3	0.545
Claude Haiku 4.5 (ontology)	28.5	21.4	25.2	14.9	21.9	77.3	0.565

Note. Accuracy values are percentages. "Mean conf." is the mean self-reported confidence (0–100) across the core set. ECE is the expected calibration error under the corresponding scoring scheme. Qwen3-32B (mechanism probe model) is reported in the main text (Table 3).

S4. Supplementary Figures

Figure S1. Four-gate diagnostic checks for the candidate_mass metric (Domain A pilot, n = 150). (a) Gate 1 — dynamic range: the norm_mass distribution spans values both below and above the uniform baseline (0) and the 2× baseline (ln 2 ≈ 0.69), so a slope is in principle estimable. (b) Gate 2 — common support: the Domain A and Domain B norm_mass distributions overlap (IoU = 0.40), supporting the cross-domain interaction test. (c) Gate 3 — mass distribution relative to the uniform baseline. (d) Gate 4 — candidate-set size |C| is not significantly associated with norm_mass (Pearson r = −0.035, p = 0.853), ruling out candidate-size bias.

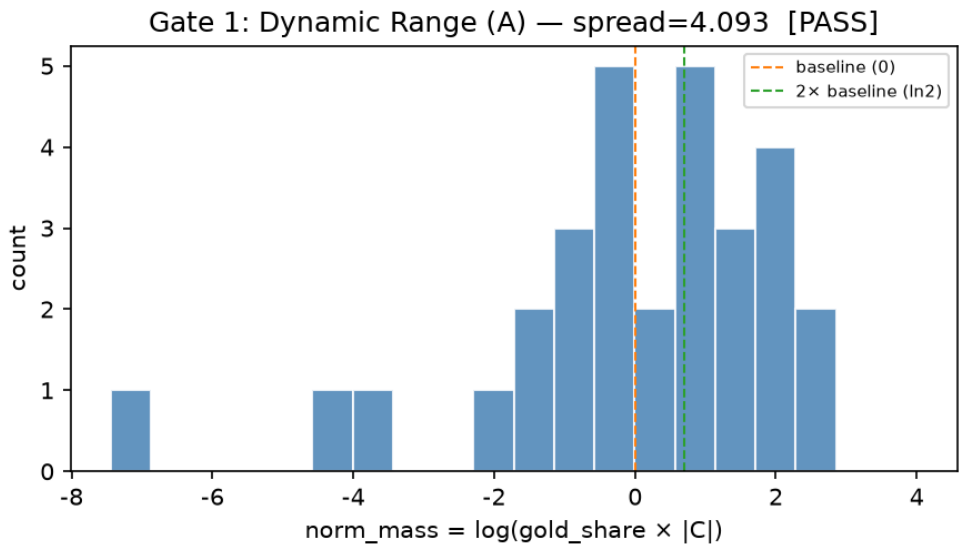

Figure S1a.

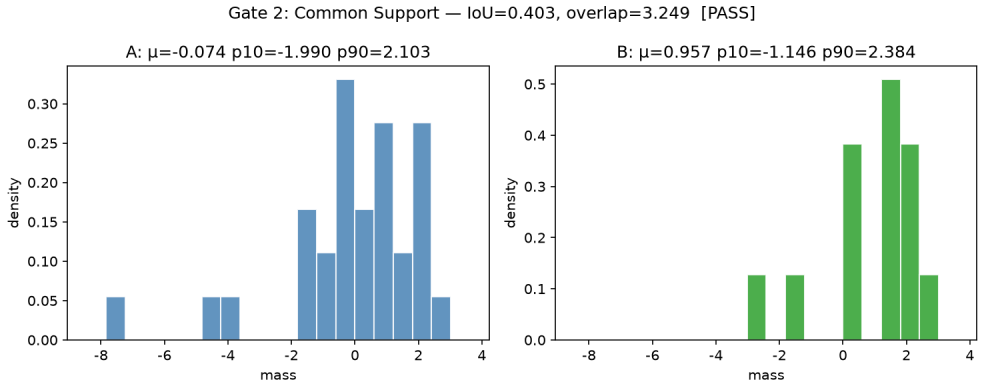

Figure S1b.

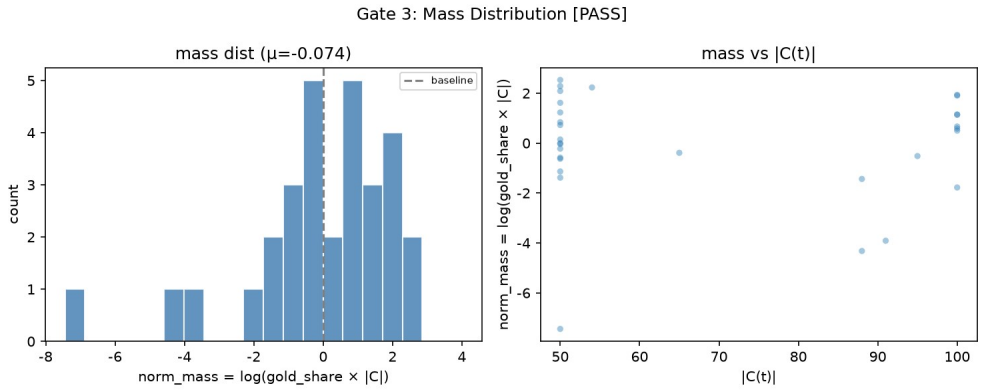

Figure S1c.

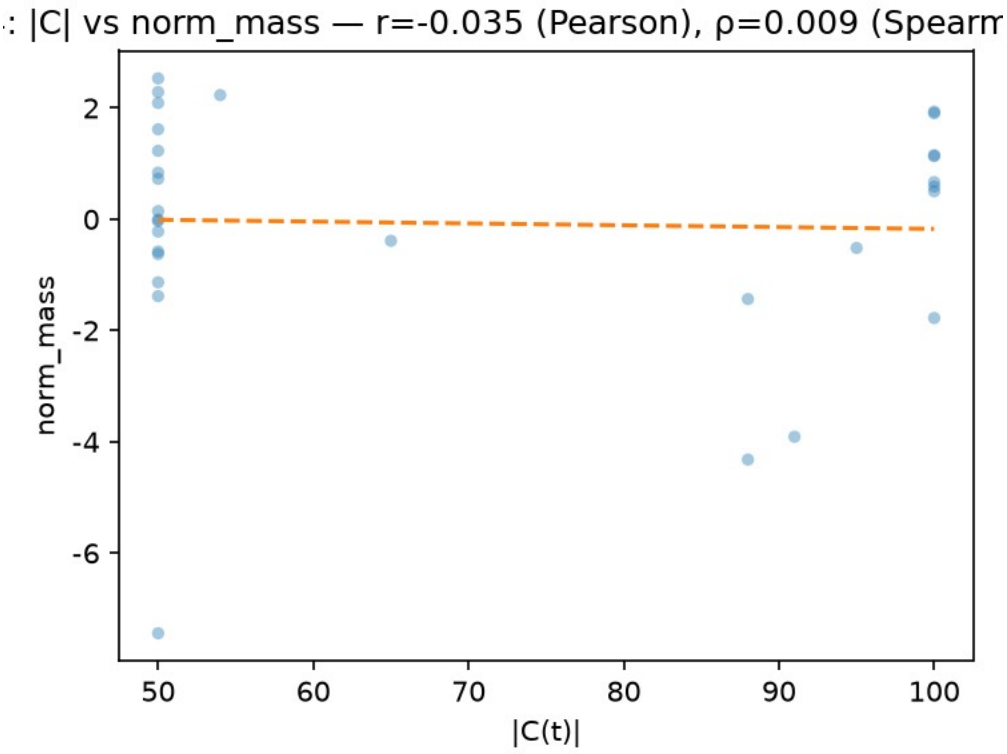

Figure S1d.

Figure S2. Reliability diagram for the three commercial models on the core set (n = 828), plotting self-reported confidence against observed accuracy. All three models are substantially overconfident at the L3 boundary (ECE = 0.506–0.623; see Table S8), consistent with the high-confidence over-alignment reported in the main text (Section 4.1).

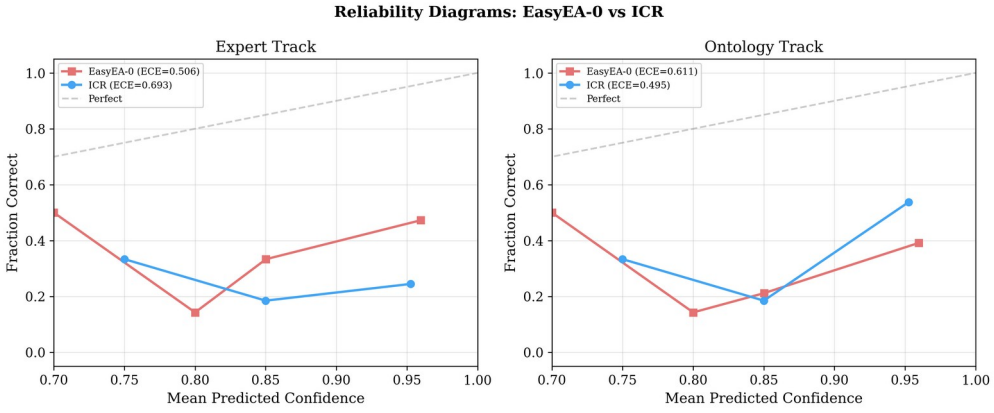

Figure S2.

S5. Main-text tables moved to Supplementary Information

The following two tables were moved from the main manuscript to the Supplementary Information to keep the main article within the Scientific Reports display-item limit while preserving their original content.

Supplementary Table S9. Definition and manipulation dimensions of the four-condition controlled probe. UC, Unconstrained; HC, Hard Constraint; RO, Reranked-Order; SC, Shuffled Constraint.

Condition	**Content Provided**	**Contains Correct Answer**	**Information Content**	**Contrastive Role**
UC	Term + definition	No	None	Lower-bound baseline
HC	UC + one correct hint	Yes	Low (sparse)	vs RO: isolates informativeness
RO	UC + reranked candidate set	Yes	High (rich)	Core mechanism condition
SC	RO candidates randomly shuffled	Yes	High (disordered)	vs RO: isolates ranking structure

Supplementary Table S10. Dataset composition and statistics.

Subset	**Domain / Content**	**Size**	**Difficulty / Type**	**Reference Standard**
Core set	Cross-lingual mathematical terms	828	L0–L3, four levels	Expert annotation
· No equivalent (L3)	Mathematics (no modern equivalent)	181	L3	Expert = no equivalent
Controlled subset A	Pre-modern Chinese mathematical terms	150	L0/L1/L2/L3 = 46/41/33/30	Expert annotation
Controlled subset B	Clinical terms → ICD	150	Equivalent 125 / Superordinate 25	Expert annotation
· Domain A candidate set	Unified candidate set C(t)	150		C
· Domain B candidate set	Unified candidate set C(t)	150		C