HARD, NOT RARE: HUMAN ITEM DIFFICULTY, NOT CONCEPT FREQUENCY, ORDERS MODEL ERRORS ACROSS PROFESSIONAL EXAMINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A common explanation for language model errors on specialised tasks is that the relevant concepts are rare in pretraining data. We test this hypothesis on individual items of the Japanese National Medical Licensing Examination, pairing exact n-gram counts from the evaluated models' own pretraining corpus with per-item human examinee accuracy, and extend the analysis to Korean licensing examinations in law, psychology, management, and chemistry, for over 4,800 items in total. The hypothesis does not hold at the item level. On 1,213 single-answer medical items, concept frequency contributes **0.30 percentage points** of cross-validated R^2 at 32B, while human item difficulty contributes **8.71 points**. In a positive control, the same counts predict per-token log-probability at $r \approx 0.76$ within fixed-length strata, so the negligible contribution on examination items cannot be attributed to measurement error. Difficulty and frequency are near-orthogonal, with a rank correlation of 0.058, and models fail hard items regardless of whether the concepts are rare. Scaling from 7B to 32B adds 25.5 percentage points on the easiest items and 6.2 on the hardest. Human item accuracy also predicts model confidence in five models from four additional families, with r between 0.223 and 0.338, and in all twenty domain–model combinations on the Korean examinations, where every 95% confidence interval excludes zero and concept frequency contributes under 2 R^2 percentage points per domain. On professional examinations, the performance gap on hard items reflects task demands that pretraining frequency does not explain.

1 INTRODUCTION

Language models now answer a large share of professional licensing examination items correctly, which makes their remaining errors informative about what limits them. A prevailing account holds that specialised knowledge lies in the long tail of web text, so the concepts that appear rarely in pretraining data are the ones models get wrong (Kandpal et al., 2023; Mallen et al., 2023). The account has a direct practical implication. If errors arise from rare concepts, more or better-targeted pretraining data should close the gap. If errors follow the intrinsic difficulty of the task independently of frequency, expanding data coverage alone will not close it.

Evidence for this account comes from factoid question answering, few-shot arithmetic, and word learning (Razeghi et al., 2022; Chang & Bergen, 2022; Mallen et al., 2023; Kandpal et al., 2023). These studies concern general-domain terms, and none compares frequency against an external measure of task difficulty. Whether the marginal frequency of the individual technical concepts an item requires governs performance on professional examinations remains untested at the item level. Such a test requires that frequencies target the specific concepts each item demands, that counts come from the evaluated models' own pretraining data (Liu et al., 2024; Soldaini et al., 2024; Team OLMo et al., 2025), and that frequency be separated from intrinsic task difficulty.

We meet these conditions on the Japanese National Medical Licensing Examination, where biomedical concept extraction and the searchable pretraining corpus of OLMo-2 permit an exact test, and

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.

then check whether the findings generalise to five further model families and to Korean licensing examinations in four professional domains, for over 4,800 items in total. The marginal frequency hypothesis does not hold at the item level. In cross-validated regressions on the medical items, concept frequency contributes 0.30 percentage points of held-out R^2 at 32B, and human item difficulty contributes 8.71 points. In a positive control, the same counts predict per-token log-probability at $r \approx 0.76$ within fixed-length strata, so the null result cannot be attributed to measurement error. Difficulty and frequency are near-orthogonal, with $r = 0.058$, and the rank correlation between human item accuracy and model confidence rises from 0.153 at 7B to 0.327 at 32B. Scaling from 7B to 32B adds 25.5 percentage points on the easiest items and 6.2 on the hardest. Human item accuracy predicts model confidence in all five additional models and in all twenty domain–model combinations on the Korean examinations, while concept frequency contributes under 2 R^2 percentage points in every setting.

Our contributions are threefold. (1) An item-level test of the marginal frequency hypothesis on professional examinations that pairs exact n-gram counts from the evaluated models' own pretraining corpus with per-item human examinee accuracy, so that difficulty and frequency enter a single regression; a token-level positive control and an answer-side replication verify the counts. (2) Evidence that the hypothesis does not hold. On medical items, human item difficulty contributes about thirty times more held-out R^2 than concept frequency, and across five model families, five professional domains, and two examination languages, model confidence tracks human item accuracy. (3) Evidence that parameter scaling concentrates gains on the items that human examinees find easiest.

2 Related Work

Frequency accounts of parametric recall. Prior work relates training frequency to model recall in two ways. Marginal frequency studies show that few-shot numerical reasoning varies with operand frequency (Razeghi et al., 2022), syntactic rule learning depends on pattern frequency (Wei et al., 2021), word frequency predicts lexical acquisition (Chang & Bergen, 2022), and entity popularity predicts when retrieval augmentation helps (Mallen et al., 2023). Co-occurrence studies count document-level question–answer entity pairs (Kandpal et al., 2023) and link subject–object co-occurrences to linear representations (Merullo et al., 2025). Recent work integrates both views to characterise parametric recall limits (Jiang et al., 2026) and connects lexical coverage to generation reliability (Zhang et al., 2025). We test the marginal frequency hypothesis at the level of individual examination items, using exact counts from the evaluated models' own pretraining data.

Memorization and parametric probing. Training frequency governs verbatim recall. Extraction rates scale with data duplication (Carlini et al., 2023; Lee et al., 2022; Tirumala et al., 2022; Biderman et al., 2023), and pretraining contents shape downstream assertions (Gekhman et al., 2024; Longpre et al., 2024). Whether pretraining frequency similarly governs performance on complex, knowledge-intensive tasks remains open. Parametric knowledge has been probed via cloze completion (Petroni et al., 2019), closed-book QA (Roberts et al., 2020), and multi-task benchmarks (Hendrycks et al., 2021; Srivastava et al., 2023). Because model probabilities can be miscalibrated and sensitive to input format (Kadavath et al., 2022; McCoy et al., 2023), we score items under a fixed format and check that results are stable across aggregation choices.

Corpus-level measurement. Computing exact training frequencies requires access to pretraining corpora. Open-weight models with documented data (Soldaini et al., 2024; Groeneveld et al., 2024; Team OLMo et al., 2025) make this possible, and suffix-array indices such as infini-gram (Liu et al., 2024) make trillion-token exact n-gram matching tractable (Elazar et al., 2024). We apply it to the fully searchable OLMo-2 pretraining mixture to obtain exact counts for every item.

Professional examinations as evaluation benchmarks. Medical licensing examinations serve as benchmarks for expert-level reasoning (Jin et al., 2020; Pal et al., 2022; Singhal et al., 2022; Nori et al., 2023). Kasai et al. (2023) curate English translations of the Japanese National Medical Licensing Examination with human examinee statistics; Liu et al. (2025) extend these evaluations to additional cohorts. Similar assessments appear in finance (Yao et al., 2025; Patel et al., 2025). Haze et al. (2023) analyse accuracy across medical subfields relative to bibliographic database document counts but use aggregate specialty-level frequencies, so difficulty and frequency cannot be separated

at the item level. Our study counts exact n-gram occurrences in the evaluated models' own pretraining corpus and measures frequency for individual items, which separates concept frequency from item difficulty within one regression.

Scaling dynamics and task difficulty. How capabilities grow with scale has motivated extensive empirical and theoretical analysis (Kaplan et al., 2020; Hoffmann et al., 2022; Wei et al., 2022; Schaeffer et al., 2023; Du et al., 2024; Lu et al., 2024). Chen et al. (2023) describe capability gains in terms of an ordered set of skills. We follow a similar approach and use human examinee accuracy to track how scaling gains distribute across levels of difficulty.

3 MEASUREMENT

We describe each item by two measurements, an external difficulty measure from human examinees and exact occurrence counts of its concepts in the evaluated models' pretraining corpus.

3.1 ITEMS AND HUMAN ITEM DIFFICULTY

We study professional licensing examinations that publish per-item examinee statistics. The primary test uses the Japanese National Medical Licensing Examination, and Korean examinations in four further domains test whether the difficulty–confidence association generalises. The Japanese examination is administered annually to graduating medical students, with each sitting comprising 400 items across six blocks. We analyse the five consecutive sittings from 2018 to 2022 curated by Kasai et al. (2023), which provide English translations alongside the exact proportion of examinees who answered each question correctly. Retaining text-only, single-answer items with verified translations and examinee statistics, evaluated across all three model scales, yields 1,226 items for accuracy analysis and 1,213 for frequency analysis; Appendix A accounts for every exclusion. Human item accuracy is the proportion of examinees who answered an item correctly in the official sitting and is independent of any model output. Higher values indicate easier items, and we use human item difficulty to refer to the ranking of items by this measure.

Korean professional examinations. We draw 1,000 items each from the Law, Psychology, and Management domains and 600 items from the Chemistry domain of the Korean Massive Multitask Language Understanding benchmark (KMMLU; Son et al., 2024). KMMLU items are sourced from Korean professional licensing examinations with per-item examinee accuracy, which serves as the external difficulty measure. All items are four-choice. We translated them into English with `deepseek-chat-v3-0324` at temperature 0, and Appendix H documents the prompt and the quality checks. Human item accuracy ranges from 1.16 to 78.57% in Law, 1.47 to 68.00% in Psychology, 4.55 to 64.15% in Management, and 7.69 to 70.73% in Chemistry.

3.2 MODELS AND SCORING PROTOCOL

We evaluate OLMo-2 Instruct at 7B, 13B, and 32B parameters (Team OLMo et al., 2025). We use the OLMo-2 family because its pretraining mixture, `olmo-mix-1124` with approximately 3.9 trillion tokens, is fully documented and searchable through suffix-array indices, so frequencies can be counted in the models' own training data. The three scales are pretrained on the same mixture, which separates the effect of scale from that of data composition.

Each item is presented with the model's chat template and scored in a single forward pass from the logits of the option-letter tokens `a`–`e`, and the prediction is the letter with the highest logit. This scoring avoids generation artefacts such as formatting variation and refusals, and it yields a continuous confidence for every item. For correlation and regression analyses we use $P(\mathrm{gold})$, the softmax-normalised probability of the correct option token over the answer choices. Binary accuracy is reported in Section 4.7.

3.3 FREQUENCY MEASUREMENT

For medical items, we extract concept spans from the question stem with the biomedical entity recogniser `en_core_sci_sm` (Neumann et al., 2019) and discard generic terms and spans

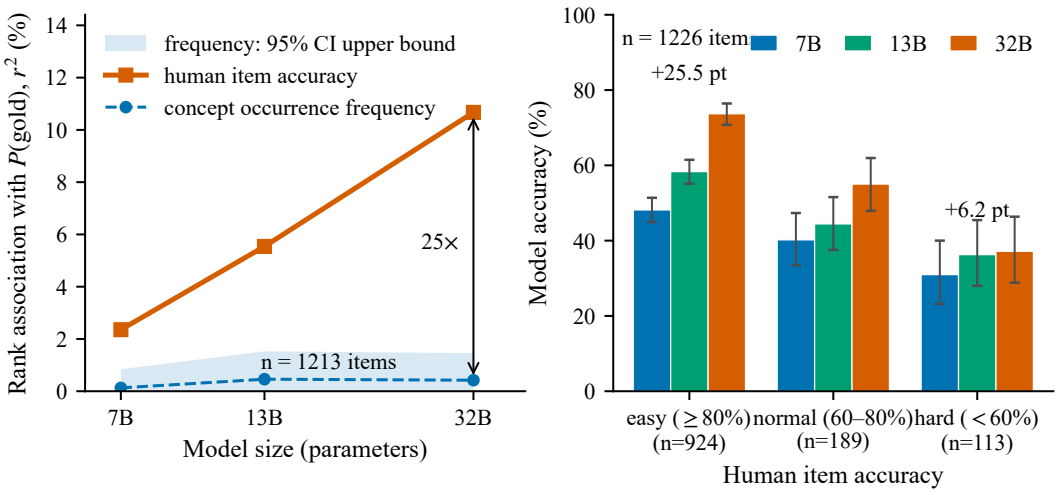

Figure 1: **Left:** Rank association (r^2) with model confidence $P(\text{gold})$ for human item difficulty and concept frequency across OLMo-2 scales. Frequency is partialled on covariates; shaded band is the 95% upper bound. **Right:** Model accuracy by human difficulty stratum (Wilson 95% CIs).

longer than five tokens. Each span is queried against the infini-gram index (Liu et al., 2024) over `olmo-mix-1124`, which returns its exact occurrence count in the corpus. Item frequency is the minimum count among an item's spans and reflects the rarest concept needed to understand the question. For KMMLU items, concept spans are noun chunks extracted with `en_core_web_lg` and counted in the same index, and item frequency is again the minimum count.

The counts have three properties relevant to the analyses that follow. Counts are exact and deterministic, introducing no sampling variance. Item frequencies range from 1 to about 6×10^8 occurrences, with 14.1% of items below 10^2 and 14.7% at or above 10^6, so analyses are not constrained by range restriction. Items containing more spans are more likely to include an infrequent concept. Span count correlates with minimum frequency at $r = -0.55$ and with stem length at $r = 0.82$, so all frequency analyses partial out span count, stem length, and human item accuracy.

3.4 Contamination and Memorisation Checks

To check for verbatim corpus contamination (Magar & Schwartz, 2022; Golchin & Surdeanu, 2024), we queried `olmo-mix-1124` with the verbatim English stems of an evenly spaced sample of 40 items. All 40 returned zero occurrences. The examinations are written in Japanese and the models read English translations (Kasai et al., 2023), which limits verbatim overlap with the pretraining corpus; KMMLU items are likewise translated from Korean. Because our dependent variables evaluate relative confidence and rankings across items rather than absolute pass rates, uniform memorisation would not induce spurious item-level associations.

3.5 Preregistered Confirmation

We established confirmation criteria for two hypotheses before examining the data: a binary contrast on a held-out item split and a low-frequency contrast re-evaluated with concepts extracted from answer options instead of stems. Both tests are reported in Section 4 alongside exploratory and sensitivity analyses.

Table 1: Held-out R^2 increment (pp) over a baseline of span count and stem length. The surface row is the strongest of eight alternatives at each scale.

Predictor added	7B	13B	32B
Human item accuracy	2.42	3.95	8.85
Strongest surface alternative	0.83	1.42	1.73

Table 2: Model accuracy (%) by human item accuracy stratum (Wilson 95% CIs). Thresholds preregistered.

Human item accuracy	n	7B	13B	32B	7B→32B
easy (≥80%)	924	48.2 [45.0, 51.4]	58.3 [55.1, 61.5]	73.7 [70.8, 76.4]	+25.5
normal (60–80%)	189	40.2 [33.5, 47.3]	44.4 [37.5, 51.6]	55.0 [47.9, 61.9]	+14.8
hard (<60%)	113	31.0 [23.2, 40.0]	36.3 [28.0, 45.5]	37.2 [28.8, 46.4]	+6.2
All	1226	45.4 [42.6, 48.1]	54.2 [51.4, 56.9]	67.5 [64.8, 70.0]	+22.1

4 RESULTS

4.1 CONFIDENCE TRACKS HUMAN DIFFICULTY

Human item accuracy predicts model confidence with rank correlations of 0.153 at 7B, 0.235 at 13B, and 0.327 at 32B (Figure 1). The progression is monotonic, and Table 3 shows that the 95% confidence intervals at 7B and 32B are disjoint.

The alignment holds out of sample. In a five-fold cross-validated linear model predicting $P(\text{gold})$ from span count, stem length, and $\log_{10}(1 + \text{count})$, adding human item accuracy increases held-out R^2 by 2.38, 3.87, and **8.71 percentage points** at 7B, 13B, and 32B, respectively.[1]

Human item accuracy is not a proxy for surface features. Table 1 compares it against eight alternative predictors, each added to a baseline of span count and stem length. The alternatives are question length, mean option length, option-length spread, negation wording, age phrasing, presence of numerals, a keyword-assigned specialty label, and all seven entered jointly. At every scale, human item accuracy contributes 2.8 to 5.1 times more held-out R^2 than the strongest surface alternative, and the specialty label contributes at most 0.10 points. Full per-predictor results appear in Appendix E.

4.2 SCALING FOLLOWS HUMAN DIFFICULTY

On the 1,226 single-answer items, overall accuracy rises from 45.4% at 7B to 54.2% at 13B and 67.5% at 32B, with 95% confidence intervals of [42.6, 48.1], [51.4, 56.9], and [64.8, 70.0]. Both steps are significant under a McNemar exact test, with $p = 2.3 \times 10^{-8}$ from 7B to 13B and $p = 2.2 \times 10^{-20}$ from 13B to 32B, and Appendix Figure 6 plots the curve. All strata remain below ceiling, so the comparisons are sensitive at every scale.

Stratifying by human item accuracy shows where the gains concentrate (Table 2, Figure 2). Scaling from 7B to 32B adds 25.5 percentage points on items that at least 80% of examinees answered correctly, 14.8 points on items between 60 and 80%, and 6.2 points on items below 60%. The largest gains fall on the items that examinees find easiest.

4.3 DIFFICULTY AND CONCEPT FREQUENCY ARE NEAR-ORTHOGONAL

Comparing difficulty and frequency requires that they measure distinct things. On these items, the rank correlation between item frequency and human item accuracy is 0.058, with a 95% confidence interval of [0.001, 0.114]. Rare concepts are not systematically the ones human candidates find hard (Figure 3). Conversely, items that most examinees answer correctly span the full frequency range from under 100 to over 10^6 occurrences. This near-orthogonality means the two variables share

[1] Table 1 reports slightly higher increments of 2.42, 3.95, and 8.85 because its baseline omits $\log_{10}(1+\text{count})$; see Appendix E.

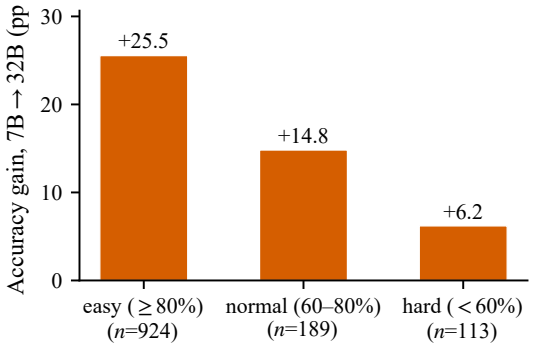

Figure 2: Accuracy gain from 7B to 32B by human difficulty stratum (n per stratum in parentheses).

Table 3: Association with model confidence $P(\text{gold})$ on 1,213 items. Top: zero-order Spearman r. Bottom: partial r controlling for span count, stem length, and human item accuracy.

Predictor	Model	r	p	95% CI	r^2
Human item accuracy	7B	0.153	<0.001	[0.098, 0.208]	2.35%
	13B	0.235	<0.001	[0.181, 0.288]	5.54%
	32B	0.327	<0.001	[0.275, 0.376]	10.67%
Concept occurrence freq.	7B	0.036	0.213	[-0.021, 0.092]	0.13%
	13B	0.068	0.018	[0.012, 0.124]	0.46%
	32B	0.065	0.024	[0.009, 0.121]	0.42%

little linear variance, and the cross-validated regressions in the next section can attribute incremental contributions with minimal collinearity.

4.4 Difficulty against concept frequency

Under the same protocol on the same items, concept frequency explains little of model confidence. After partialling out span count, stem length, and human item accuracy, the partial rank correlation between concept frequency and $P(\text{gold})$ across 1,213 items is 0.036 at 7B, 0.068 at 13B, and 0.065 at 32B, with p values of 0.21, 0.018, and 0.024 (Table 3). The widest 95% upper bound across all three scales is 0.124, stable across all aggregation metrics in Appendix B.

The gap is equally clear in predictive terms. In a five-fold cross-validated regression controlling for span count, stem length, and human item accuracy, adding $\log_{10}(1 + \text{count})$ raises held-out R^2 by **0.30 percentage points** at 32B, and by 0.04 and 0.32 points at 7B and 13B. Adding human item accuracy to a model that already contains frequency yields **8.71 percentage points** at 32B. On the same items under the same regression, difficulty provides roughly thirty times the predictive contribution of frequency.

Corrections for multiple comparisons and alternative aggregations point in the same direction. Under a Bonferroni-corrected threshold of $\alpha = 0.017$ across the three scales, none of the three partial correlations reaches statistical significance. Replacing the minimum with the median, geometric mean, or maximum count across spans yields null associations at all scales. A preregistered binary contrast on a held-out split finds no frequency difference, with $p = 0.64$ at 32B and $p > 0.6$ at 7B and 13B.

4.5 Token-level positive control

A positive control verifies that the counts are sensitive at the level where frequency is expected to operate. For each of 4,138 distinct concept spans, we compute the mean per-token log-probability of the span conditioned on a beginning-of-sequence token. Table 4 reports spans of two to seven tokens, and Appendix F describes the remaining spans. Because frequent spans tend to be shorter

Table 4: Spearman r between $\log_{10}$ count and per-token log-probability by token count (span counts in parentheses).

Tokens in span	2 (1256)	3 (1236)	4 (637)	5 (341)	6 (187)	7 (81)
32B	0.79	0.79	0.80	0.80	0.81	0.82
7B	0.75	0.77	0.80	0.79	0.81	0.84

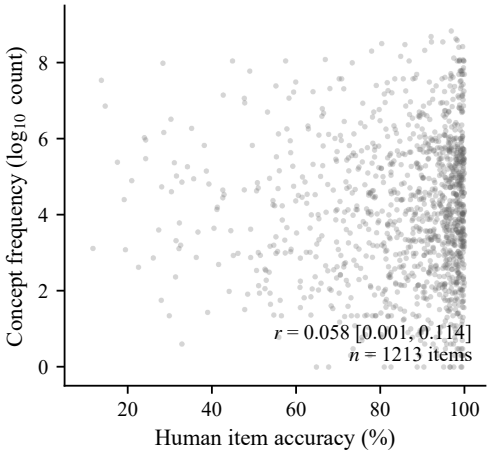

Figure 3: Human item accuracy against concept frequency for 1,213 items, $r = 0.058$.

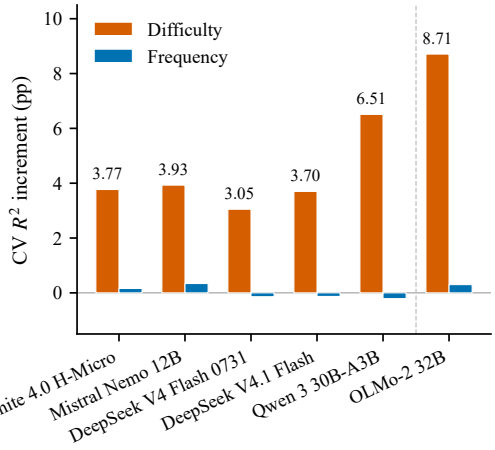

Figure 4: Cross-validated R^2 increments for difficulty and frequency across five families. OLMo-2 32B, right of the dashed line, is shown for reference.

and longer spans score higher per token from internal-prefix predictions, we stratify by token count before correlating with $\log_{10}$ occurrence count.

Within every stratum from 2 to 7 tokens, the Spearman correlation between occurrence count and mean per-token log-probability exceeds $+0.75$, with no systematic drop at longer spans (Table 4). The size-weighted correlation across all strata is $+0.76$ at 32B and $+0.72$ at 7B.

The counts that track token-level log-probability this closely have a partial correlation of at most 0.068 with examination confidence on the items that contain the same spans.

4.6 Answer-side replication

An apparent low-frequency effect in question stems does not replicate when concepts are extracted from answer options. Binning items by frequency decade shows that $P(\text{gold})$ does not follow a simple monotonic curve (Appendix Figure 7). For items whose rarest stem concept occurs fewer than 100 times, 32B model confidence is 0.166 lower than for items in the 10^4 to 10^5 decade, with $p = 4.1 \times 10^{-5}$. The contrast remains after matching on span count, where it is 0.178, and on stem length, where it is 0.151. Human examinees answer the two subsets within 0.5 percentage points of each other, the contrast appears in all five annual sittings, and it persists in both vignette and non-vignette items, at 0.182 and 0.214.

To test whether this effect reflects conceptual rarity or extraction location, we applied a preregistered answer-side replication, extracting concepts from the correct answer option instead of the question stem. Stem-derived concepts capture clinical descriptions; answer-derived concepts capture specific diagnoses and treatments. Both extractions use the same pretraining index and counting algorithms. On the answer side, the gap narrows to 0.038 at 32B, 0.012 at 13B, and 0.042 at 7B, with $p > 0.4$ at every scale. The 95% upper bound at 32B is 0.156, below the stem-side estimate of 0.166.

Covariate matching, subgroup stratification, and format controls all supported the stem-side effect, yet it did not replicate on concepts extracted from answer options. Corpus statistics aggregated over extracted spans should therefore be checked against a second extraction source.

Table 5: Cross-family results on the medical items. The number of items with valid option probabilities ranges from 1,123 for Qwen 3 to 1,214. Frequency uses `olmo-mix-1124` counts throughout.

Model	r [95% CI]	CV R^2 increment (pp)	
		Difficulty	Frequency
DeepSeek V4.1 Flash	0.268 [0.215, 0.319]	3.70	−0.13
DeepSeek V4 Flash 0731	0.259 [0.206, 0.311]	3.05	−0.14
Mistral Nemo 12B	0.229 [0.174, 0.283]	3.93	0.34
Qwen 3 30B-A3B	0.338 [0.286, 0.389]	6.51	−0.21
Granite 4.0 H-Micro	0.223 [0.168, 0.275]	3.77	0.16
OLMo-2 32B (reference)	0.327 [0.275, 0.376]	8.71	0.30

4.7 Robustness and sensitivity analyses

The primary findings hold under alternative measures. Substituting binary accuracy for $P(\text{gold})$ preserves the directional patterns. Human item accuracy correlates with correctness at 0.141, 0.200, and 0.263 at 7B, 13B, and 32B, and the partial correlation of concept frequency falls to 0.009, 0.037, and 0.046. To rule out priors over option letters, we also use the logit margin between the correct answer and the top distractor. This margin-based metric yields associations of 0.144, 0.230, and 0.326 for human difficulty and 0.031, 0.065, and 0.080 for concept frequency, close to those under $P(\text{gold})$. Stratifying by examination year gives the same direction for both predictors in all five sittings. Complete permutation tests, order-statistic diagnostics, and alternative aggregation grids appear in Appendix B.

Cross-family results. To test whether the association generalises across model families, we evaluated five additional models from four families using the same log-probability scoring on the same item pool (Table 5, Figure 4). Human item accuracy predicts model confidence with r between 0.223 and 0.338 across all five models. In every cross-validated regression, difficulty contributes 3.05 to 6.51 percentage points of held-out R^2 while concept frequency contributes at most 0.34 points. Because the pretraining corpora of these models are not public, their frequency counts use `olmo-mix-1124` as a proxy. The frequency result therefore rests on OLMo-2, where the corpus is matched exactly, and the cross-family test shows that the difficulty–confidence association holds across architectures and training corpora. The models cover dense, mixture-of-experts, and hybrid state-space architectures and range in accuracy from 47.7% for Granite to 78.7% for DeepSeek V4.1, against 67.5% for OLMo-2 32B.

Cross-domain and cross-language results. To test whether the association generalises beyond medicine and beyond one examination language, we evaluated OLMo-2 at three scales and four cross-family models on 3,600 items from the Law, Psychology, Management, and Chemistry domains of KMMLU. Across all twenty domain–model combinations, the Spearman correlation between human item accuracy and model confidence is positive, with r from 0.068 to 0.217 and every 95% confidence interval excluding zero (Table 6, Figure 5). The association holds in domains whose subject matter does not overlap with medicine, from legal reasoning to quantitative problem solving in chemistry, and across the two examination languages, Japanese and Korean. Appendix Table 9 shows that OLMo-2 accuracy rises monotonically from 7B to 32B in every domain, consistent with the scaling pattern in the medical domain. Under the same cross-validated regression, concept frequency contributes under 2 R^2 percentage points in each KMMLU domain, ranging from -0.48 to $+1.62$ at 32B, and difficulty contributes $+0.55$ to $+3.26$ points.

5 Discussion

Frequency counts predict the log-probability of individual concept spans but not model confidence on the items that contain them. One interpretation is that professional examination items are difficult because of how concepts must be combined within a clinical, legal, or scientific scenario, and familiarity with each concept is not sufficient to combine them correctly. The scaling results are consistent with this interpretation. Larger models gain most on items that examinees find easy and

Table 6: Cross-domain results on KMMLU. Spearman r between human item accuracy and model confidence $P(\text{gold})$ (95% Fisher CIs). All confidence intervals exclude zero.

Domain (n)	OLMo-2 32B	DeepSeek V4.1	Mistral Nemo 12B	Qwen 3 30B-A3B	Granite 4.0 H-Micro
Law (1,000)	0.176 [0.115, 0.235]	0.194 [0.133, 0.253]	0.103 [0.042, 0.164]	0.217 [0.157, 0.276]	0.136 [0.074, 0.196]
Psychology (1,000)	0.140 [0.079, 0.201]	0.104 [0.042, 0.165]	0.127 [0.065, 0.187]	0.122 [0.061, 0.183]	0.144 [0.082, 0.204]
Management (1,000)	0.136 [0.075, 0.196]	0.144 [0.083, 0.204]	0.068 [0.006, 0.130]	0.177 [0.117, 0.237]	0.094 [0.032, 0.155]
Chemistry (600)	0.136 [0.056, 0.214]	0.186 [0.106, 0.263]	0.088 [0.008, 0.166]	0.184 [0.106, 0.260]	0.156 [0.077, 0.233]
IgakuQA (ref.)	0.327 [0.275, 0.376]	0.268 [0.215, 0.319]	0.229 [0.174, 0.283]	0.338 [0.286, 0.389]	0.223 [0.168, 0.275]

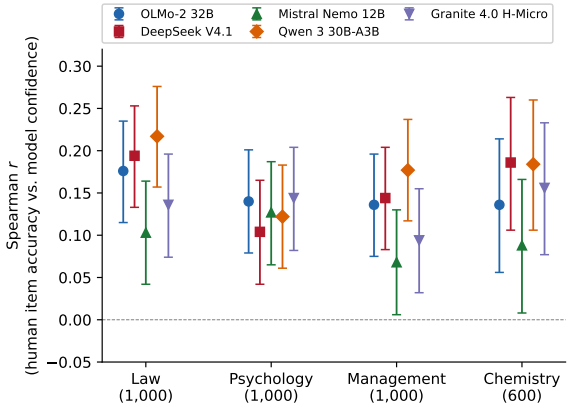

Figure 5: Cross-domain results on KMMLU. Spearman r between human item accuracy and model confidence across five models and four professional domains (95% Fisher CIs). All twenty confidence intervals exclude zero.

least on items that examinees find hard, even though easy and hard items do not differ systematically in concept frequency. Our test addresses the marginal frequency of individual concepts; whether document-level co-occurrence (Kandpal et al., 2023; Merullo et al., 2025) explains additional variance remains open.

The difficulty–confidence association holds across five model families and five professional domains. Despite differing architectures and training mixtures, each model tends to assign higher confidence to items that more examinees answer correctly. This agreement suggests that the difficulty ordering is a property of the items themselves and does not depend on the training pipeline of a particular model.

6 Conclusion

We test whether language model errors on professional examinations reflect data scarcity or task difficulty. Pairing exact pretraining counts with per-item examinee statistics on medical licensing items, we find that concept frequency contributes 0.30 R^2 percentage points at 32B against 8.71 for human difficulty, with scaling gains concentrating where human performance is highest. The difficulty–confidence association replicates across two languages, five model families, and twenty domain–model combinations. Hard items remain hard across architectures and scales, and concept frequency does not explain this difficulty, which suggests that broader pretraining coverage of individual concepts is unlikely to close the gap on these items.

AI use statement

We used a large language model for two purposes. We translated the 3,600 Korean KMMLU items into English with `deepseek-chat-v3-0324` (Section 3); Appendix H documents the prompt and the quality checks, and the translations are released with the supplementary material. We also used a language model to edit the manuscript for readability. The authors reviewed all AI-assisted output and take full responsibility for the content of this paper.

Reproducibility statement

Section 3 specifies the item pool, the scoring protocol, and the frequency measure, including the entity recogniser, the span filters, the index queried, and the aggregation rule. Appendix A accounts for every item from the 2,000 released to the 1,213 entering the frequency analyses. Appendix B reports the aggregation grid, the permutation test, the order-statistic diagnostic, and the stratification by sitting, and Appendix C the simpler frequency measures that we tested and discarded. The medical items, their English translations, and the per-item examinee statistics are public (Kasai et al., 2023), as are the KMMLU items and their statistics (Son et al., 2024), the pretraining index, and the evaluated models (Liu et al., 2024; Team OLMo et al., 2025). Our English translations of the KMMLU items are included in the supplementary material. The counts are exact and deterministic, so the frequency measurements are reproducible as specified. Code for the measurement pipeline and for every analysis reported here, together with the per-item derived data, is provided as anonymous supplementary material and will be released publicly upon publication.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Tyler A. Chang and Benjamin K. Bergen. Word acquisition in neural language models. *Trans. Assoc. Comput. Linguistics*, 10:1–16, 2022. doi: 10.1162/tacl_a_00444.

Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! A data-driven skills framework for understanding and training language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, 2023.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024*, 2024.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data? In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pp. 7765–7784. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.444.

Shahriar Golchin and Mihai Surdeanu. Time travel in llms: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024.

Dirk Groeneveld, Iz Beltagy, Evan Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, pp. 15789–15809. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.841.

Tatsuya Haze, Rina Kawano, Hiromichi Takase, Shota Suzuki, Nobuhito Hirawa, and Kouichi Tamura. Influence on the accuracy in ChatGPT: Differences in the amount of information per medical field. *International Journal of Medical Informatics*, 180:105283, 2023. doi: 10.1016/j.ijmedinf.2023.105283.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Changhao Jiang, Ming Zhang, Yifei Cao, Junjie Ye, Xiaoran Fan, Shihan Dou, Zhiheng Xi, Jiajun Sun, Yi Dong, Yujiong Shen, Jingqi Tong, Baoyu Fan, Tao Gui, Qi Zhang, and Xuanjing Huang. Beyond scaling: Measuring and predicting the upper bound of knowledge retention in language model pre-training. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026*, pp. 26377–26396. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1214.

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. *arXiv preprint arXiv:2009.13081*, 2020.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating GPT-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*, 2023.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pp. 8424–8445. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.577.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024.

Junyu Liu, Lawrence K. Q. Yan, Tianyang Wang, Qian Niu, Momoko Nagai-Tanima, and Tomoki Aoyama. Kokushimd-10: Benchmark for evaluating large language models on ten japanese national healthcare licensing examinations. In *AI for Clinical Applications - First International Workshops, Agentic AI 2025, CREATE 2025, and Clinical MLLMs 2025, Held in Conjunction with MICCAI 2025*, volume 16147 of *Lecture Notes in Computer Science*, pp. 300–309. Springer, 2025. doi: 10.1007/978-3-032-06004-4_30.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3245–3276. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.179.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pp. 5098–5139. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.279.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022*, pp. 157–165. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.18.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*, 2023.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. On linear representations and pretraining data frequency in language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net, 2025.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019*, pp. 319–327. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-5034.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

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

Jaisal Patel, Yunzhe Chen, Kaiwen He, Keyi Wang, David Li, Kairong Xiao, and Xiao-Yang Liu. Reasoning models ace the CFA exams. *arXiv preprint arXiv:2512.08270*, 2025.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pp. 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-1250.

Yasaman Razeghi, Robert L. Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.59.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pp. 5418–5426. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.437.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Raghavi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pp. 15725–15788. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.840.

Guijin Son, Hanwool Kim, Hyunwoo Moon, et al. KMMLU: Measuring massive multitask language understanding in Korean. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3131–3163, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cèsar Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar,

Fernando Martínez-Plumed, Francesca Happé, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse H. Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory W. Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, María José Ramírez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael I. Ivanitskiy, Michael Starritt, Michael Strube, Michal Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T., Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Milkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay V. Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.*, 2023, 2023.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri,

Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pp. 932–948. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.72.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022.

Xuan Yao, Qianteng Wang, Xinbo Liu, and Ke-Wei Huang. Evaluating large language models for financial reasoning: A CFA-based benchmark study. *arXiv preprint arXiv:2509.04468*, 2025.

Shuo Zhang, Fabrizio Gotti, Fengran Mo, and Jian-Yun Nie. Measuring the impact of lexical training data coverage on hallucination detection in large language models. *arXiv preprint arXiv:2511.17946*, 2025.

A ITEM ACCOUNTING

Step	Items
Five sittings, 2018–2022 (six blocks each)	2,000
fewer items with images, without an English translation, or without an examinee statistic	−538
Eligible item pool	**1,462**
fewer items whose released English stem is not a rendering of the Japanese	−4
Entering frequency measurement	1,458
fewer items yielding no countable medical concept span	−11
With a frequency value and an examinee statistic	1,447
fewer multi-answer items and items missing a score at any scale	−234
Frequency analyses	**1,213**
Single-answer items scored at all three scales (accuracy analyses)	1,226

All analyses start from the eligible pool of 1,462 items. Four items whose released English field does not render the stem are detected by string match and removed before span extraction. Accuracy analyses (Section 4.2) do not need a frequency value and use 1,226 items, while frequency analyses use the 1,213 items that also have one. The thirteen items in the first set but not the second comprise eleven items without a countable medical concept span and two of the four items removed for their English rendering.

B ROBUSTNESS

Aggregation grid. Partial rank correlations between item frequency and $P(\mathrm{gold})$, controlling span count, stem length, and human item accuracy, for four aggregations of an item's span counts:

Asterisks mark nominal $p < 0.05$; two of twelve estimates reach it, and none clears a Bonferroni threshold over the grid. The widest 95% upper bound across all twelve is 0.124.

Aggregation	7B	13B	32B
Minimum	0.036	0.068^*	0.065^*
Median	0.043	0.042	0.016
Geometric mean	0.033	0.048	0.029
Maximum	0.016	0.022	-0.018

Order-statistic diagnostic. Span count correlates with the minimum aggregation at $r = -0.55$ and with stem length at $r = 0.82$, while stem length correlates with human item accuracy at $r = -0.02$. The order statistic therefore affects the predictor and not the outcome, which is why span count and stem length are partialled out throughout. The maximum aggregation is not subject to the order statistic, and under it the association is null at every scale.

Permutation test. Shuffling item frequency across items 10,000 times while holding the outcome fixed gives a null distribution of $|r|$ with median 0.020 and 95th percentile 0.056 at 32B, against an observed 0.087 before covariate adjustment, giving an empirical $p = 0.003$.

Stratification by sitting. At 32B, zero-order Spearman correlations with $P(\text{gold})$ by year are 0.082, 0.060, 0.049, 0.080, and 0.154 for frequency and 0.387, 0.284, 0.292, 0.348, and 0.308 for human item accuracy. Both predictors are positive in all five sittings, and human item accuracy exceeds frequency in each by a factor of two to six.

Translation quality. Fifty-eight items, 5.0% of the pool before filtering, have English fields in which at least one option is left unrendered, typically options referring to circled numerals in the original. These items are excluded. Recomputing the scale comparison with and without them moves accuracies by less than 1.5 points at every scale and leaves both adjacent contrasts significant.

C SIMPLER FREQUENCY MEASURES

Three simpler ways of turning an item into a frequency value do not yield a valid predictor.

Whole option text as an n-gram. In a probe of 298 answer options, submitting each option verbatim yields a zero count for 44% of them. Zeros concentrate in options of four or more tokens, with 124 zero and 24 non-zero counts, against 7 zero and 143 non-zero counts for options of one to three tokens, because many options are full clauses and not terms. The resulting variable tracks the phrasing of a translation more than the presence of a concept.

Rarest sub-n-gram of one to three tokens. Enumerating short sub-spans and taking the rarest non-zero count removes the zeros but selects grammatical fragments such as prepositional openings and modal fragments, because rarity attaches to awkward token boundaries. The measurement problem moves from the clause level to the fragment level rather than being solved.

Controlled-vocabulary filtering. Restricting candidate spans to exact MeSH descriptor matches rejects all six fragments in a hand-checked probe but accepts only three of eight genuine medical concepts, since exam-level phrasings such as "osmotic laxative" and "acute intestinal necrosis" are not preferred descriptors. At that recall the variable would report the shape of the vocabulary as much as the content of the corpus.

Biomedical span extraction (Section 3) is subject to none of these failures and leaves only 0.8% of items without a countable concept.

D A SECOND MODEL FAMILY

Pythia is the only open family we identified whose pretraining corpus is indexed on the same infrastructure as OLMo-mix, so the same frequency measure can be computed for it. Across the 1,444 items with counts in both indices, Pile and OLMo-mix rank concepts almost identically, with a Spearman correlation of 0.94, but differ sharply in level. The median item count is 1,868 in the Pile and 11,117 in OLMo-mix, and 22.5% of items fall below 10^2 in the Pile against 14.3% in OLMo-mix.

Performance does not transfer. Pythia-6.9B scores 17.5% on 338 single-answer items against a chance level of 20%, with a one-sided binomial $p = 0.89$. Of its predictions, 86.4% collapse onto a single option letter, and accuracy does not vary across human-difficulty strata, at 17.1%, 16.7%, and 21.1% on easy, normal, and hard items. Because Pythia is a base model without a chat template, we scored it by completion rather than by the chat-formatted protocol used for OLMo-2, and we checked that the low accuracy is not an artefact of that choice. We tested four standard base-model formats: an instruction preamble with `Answer:`, a bare stem, a parenthesised letter cue, and a one-shot example. These yield 18.4%, 18.0%, 18.4%, and 19.2%, all indistinguishable from chance and all showing the same collapse. We therefore did not extend the family to larger scales.

E HUMAN DIFFICULTY AGAINST SURFACE PREDICTORS

To test whether human item accuracy is merely a proxy for a surface cue, we compare it against eight alternatives under the protocol of Section 4.4: question length, mean option length, option-length spread, negation wording, age phrasing, presence of numerals, a keyword-assigned specialty label, and all seven entered jointly. Every candidate predictor is added to the same baseline of span count and stem length, which omits $\log_{10}(1 + \text{count})$; the increments below therefore sit on a slightly different footing from the 2.38, 3.87, and 8.71 points of Section 4.1. Held-out R^2 increments over the two-covariate baseline, in percentage points:

Table 7: Held-out R^2 increment (pp) per predictor added to a baseline of span count and stem length ($n = 1{,}213$). This baseline omits $\log_{10}(1 + \text{count})$; see Section 4.1 for the three-covariate comparison.

Predictor added	7B	13B	32B
Human item accuracy	2.42	3.95	8.85
Negation cue	0.83	0.13	−0.22
Option-length spread	−0.06	0.99	1.73
Mean option length	0.06	1.12	1.13
Question length	0.06	0.05	−0.07
Concept frequency	0.07	0.39	0.43
Age phrasing	−0.10	−0.12	−0.03
Numerals	−0.09	−0.15	−0.28
Specialty label	−0.78	0.10	−0.40
All surface combined	−0.56	1.42	0.80

Human item accuracy exceeds every surface alternative at every scale. The strongest surface predictor is the negation cue at 7B, the joint block at 13B, and option-length spread at 32B, so no surface cue gives a stable signal across scales. The specialty label, assigned by keyword match, contributes at most 0.10 points at any scale.

F FREQUENCY AND TOKEN LOG-PROBABILITY

For each of the 4,138 distinct spans we compute the mean per-token log-probability of the span under the model, conditioned only on a beginning-of-sequence token. Stratified analyses use the 4,088 spans of one to seven tokens, and the 50 longer spans are not stratified. The raw correlation with $\log_{10}$ count is negative, at -0.31 for 32B and -0.35 for 7B, because of span length. Frequent spans are shorter, with a Spearman correlation of -0.68 between count and token length, and longer spans score higher on a per-token mean because their internal tokens follow from their own prefixes, with a correlation between token length and mean log-probability of $+0.88$ at 7B and $+0.85$ at 32B. Together, the two confounds reverse the sign of the raw association.

Within strata of fixed token count the relationship is strong, positive, and stable (Table 8).

Weighting all strata by size, including the single-token stratum, gives $+0.76$ at 32B and $+0.72$ at 7B; restricting to the multi-token strata shown in Table 4 gives $+0.79$ at 32B and $+0.77$ at 7B.

Table 8: Spearman r between $\log_{10}$ count and per-token log-probability by token count. The 350 single-token spans are listed separately because they have no preceding context and a narrow frequency range.

Tokens in span	1	2	3	4	5	6	7
Spans	350	1256	1236	637	341	187	81
Spearman, 32B	0.43	0.79	0.79	0.80	0.80	0.81	0.82
Spearman, 7B	0.19	0.75	0.77	0.80	0.79	0.81	0.84

G KMMLU SCALING DETAILS

Table 9: OLMo-2 scaling on KMMLU. Accuracy (%) and Spearman r between human item accuracy and model confidence $P(\text{gold})$ (95% Fisher CIs). Accuracy rises monotonically from 7B to 32B in every domain.

	Accuracy (%)			r [95% CI]		
Domain (n)	7B	13B	32B	7B	13B	32B
Law (1,000)	38.3	39.5	47.3	0.135 [0.074, 0.196]	0.080 [0.018, 0.141]	0.176 [0.115, 0.235]
Psychology (1,000)	37.3	42.1	49.0	0.109 [0.047, 0.170]	0.123 [0.061, 0.183]	0.140 [0.079, 0.201]
Management (1,000)	44.9	50.6	59.9	0.041 [−0.021, 0.103]	0.131 [0.069, 0.191]	0.136 [0.075, 0.196]
Chemistry (600)	39.2	44.3	57.7	0.042 [−0.038, 0.121]	0.084 [0.004, 0.163]	0.136 [0.056, 0.214]

H KMMLU TRANSLATION

Each of the 3,600 KMMLU items was translated from Korean into English in a separate request to `deepseek-chat-v3-0324` at temperature 0. The user message contained the Korean question and its answer options as JSON, and the system prompt was the following.

> You are a professional translator. Translate the following Korean professional exam question and its answer choices into English. Preserve all technical terminology accurately. Output ONLY valid JSON with keys “question” and “choices” (a list of strings). Do not add explanations.

Quality control was automatic. Every item returned a parseable translation, all 3,600 translations keep the number of answer options of the original, and none has an empty question or option. Seven items retain Korean characters, and in each case the retained text is appropriate: one legal term is followed by its Korean original in parentheses, three items name fictional Korean companies, and three chemistry items use the Korean interpunct as the multiplication dot in a unit. The English text is a median of 2.62 times as long as the Korean in characters, with the 1st and 99th percentiles at 1.26 and 4.11. The Korean originals, the English translations, the translation script, and the checks are included in the supplementary material.

Human item accuracy is measured on the Korean originals while models read the English translations, the same arrangement as for the medical items. Translation errors that are unrelated to item difficulty add noise to model confidence without changing human item accuracy, which tends to attenuate the correlation between the two. The positive associations in all twenty domain–model combinations of Table 6 are therefore conservative with respect to such errors.

I SUPPLEMENTARY FIGURES

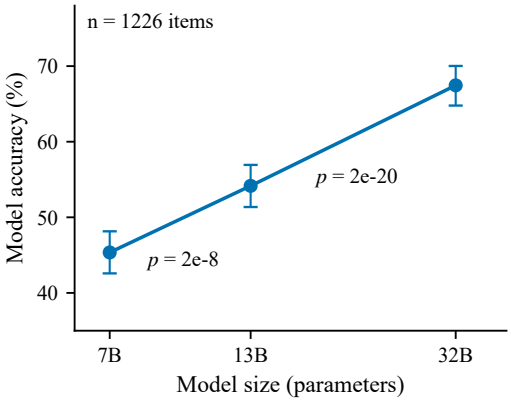

Figure 6: Model accuracy by parameter scale (Wilson 95% CIs, McNemar p).

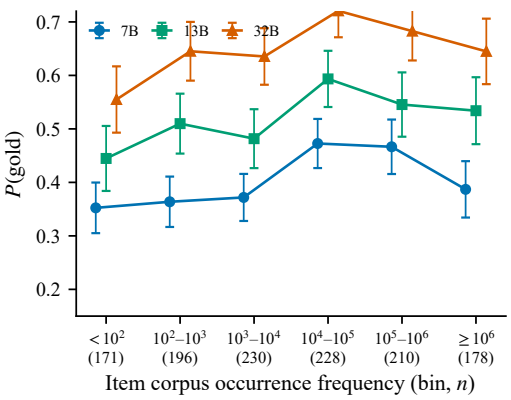

Figure 7: Mean model confidence $P(\text{gold})$ across item frequency decades with 95% confidence intervals (sample counts in parentheses).