DERIVING SCALING LAWS FOR LLM BEHAVIOR VIA RATIONAL USE OF REPRESENTATIONAL RESOURCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling laws forecast language model loss from parameter count and dataset size, but they underdetermine behavior (language models at a given loss can behave arbitrarily differently), and thus cannot model why, or when, behaviors emerge in models. We propose, instead, to rationalize language models not as suboptimal solutions to an objective specified by the data, but as the *optimal* solutions to an objective jointly defined by the data and the resources available (parameter count, dataset size, etc.), effectively as zero-loss models of a distribution compressed by finite resources. By doing so we connect this approach to the theory of lossy compression. We design a general method to solve for a data-generating process' optimal solutions and show that, in synthetic language modeling tasks, models are generally optimal to within 0.02 nats/token. While these optima are intractable to compute exactly for large corpora, we present a scalable approximate method and analyze in-the-wild language model scientific knowledge as optimal with respect to compressed research taxonomies. Across scales, we are able to predict which facts models acquire, and, for unacquired facts, we predict the evolution of hallucinations with scale, providing empirical support for the theory.

1 INTRODUCTION

Recent progress in AI has been driven by scale: larger models and more data. As models get larger they come closer to capturing the data-generating process behind language, and with it the full productive capacity of the human intelligence behind the data. In practice, however, with finite resources, the data-generating process itself is never attained. When resources bind, loss underdetermines behavior: language models with matched loss can differ wildly in downstream behavior (Liu et al., 2023; Abbe et al., 2024). The language models we produce are approximations, with surprising behaviors that can appear abruptly (Ganguli et al., 2022; Wei et al., 2022) and disappear with scale (McKenzie et al., 2023), only sometimes to recover with even more (Wei et al., 2023).

Though these behaviors are surprising, they are also predictable. Abilities emerge at common loss across models (Du et al., 2024), and models of different sizes perform alike in context (Xia et al., 2023); downstream performance is predictable across model families in aggregate or through intermediate quantities (Ruan et al., 2024; Gadre et al., 2025; Owen, 2024; Bhagia et al., 2025); memorization tracks frequency (Razeghi et al., 2022; Kandpal et al., 2023; Mallen et al., 2023; Carlini et al., 2023); and circuits, grammar and n-gram statistics form at predictable stages of training (Olsson et al., 2022; Chang & Bergen, 2022; Choshen et al., 2022; Bietti et al., 2023; Edelman et al., 2024; Varre et al., 2025; Yüksel et al., 2026). These regularities hold across seed, size, and family, suggesting the principles behind this behavior are general. The clearest form of this predictability is in scaling laws: as language models grow, their loss falls along power laws regular enough to fix model and data sizes before training (e.g., Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022; OpenAI, 2023).

These results are encouraging because they motivate the search for general principles underlying these regularities. They promise that language model behavior, and not just loss, may be predictable. In this paper, we argue that these regularities can be explained by models making efficient use of their representational resources. This perspective invites interpretation of the language-modeling loss as a measure of compression: the log loss of a text is its code length under the model (Shannon, 1948; Delétang et al., 2024), and a model is a (perhaps lossy) projection of the true distribution.

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.

Under this reading, trained models are not failed attempts to attain the data-generating process, but rather optimal approximations to it, given the constraints on the resources available.

This approach allows us to use methods from information theory, specifically rate–distortion theory, to quantitatively predict model behavior as a function of scale. In synthetic tasks we predict power-law scaling and cases where language models acquire parts of the data-generating process exactly, as well as cases where they impose structure that isn't present in the generating process. At scale, our methods for approximately characterizing optimal compressions predict acquisition and lawful hallucination in language model factual recall across four model families, 70M–33B parameters, and 0.3T–36T training tokens, in 26 large language models and 97 checkpoints.

2 Related Work

Scaling laws. Loss falls as a power law in parameters, data and compute (Kaplan et al., 2020; Hoffmann et al., 2022). Dębowski (2026) shows that neural scaling follows from Zipf's law under broad assumptions, by way of Heaps' law and Hilberg's hypothesis that the information adjacent blocks of text share grows as a power of their length. Further, information theory yields a compute-optimal relation between data and model size that is linear up to logarithmic factors (Jeon & Van Roy, 2024).

Description length and implicit simplicity bias. Accounts of learning as compression ground the available representational resources in different ways: Minimum description length selects the model that minimizes its own description plus that of the data (Rissanen, 1978). Bits-back coding sends a model drawn from Q to a receiver who holds the prior π at cost $\mathrm{KL}(Q\|\pi)$ nats (Hinton & van Camp, 1993), where $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence. PAC-Bayes bounds built on the same machinery also apply to large language models (Lotfi et al., 2024).

Kolmogorov's structure function gives each sequence's best model at every complexity (Vereshchagin & Vitányi, 2004). Rate–distortion theory (Shannon, 1959) solves relatives of the budgeted problem of § 3. The information bottleneck, for example, adds a relevance variable (Tishby et al., 1999); it is the problem in § 3 without its description term. Applied to sequences, the information bottleneck yields optimal predictive models for a given complexity (Still et al., 2010). Along its solution paths, clusters split at critical temperatures (Rose, 1998). In § 3 we model the shared structure of tasks with prototypes and add a description cost per prototype, which makes each split of a prototype away from the default first-order (Prediction 2).

Trained networks approach information-theoretic bounds in other settings. Pretraining drives hidden representations toward the bottleneck bound, and proximity to it predicts performance (Conklin et al., 2026). Networks prefer low-complexity solutions without an explicit penalty (Goldblum et al., 2024). Information bottleneck accounts of deep learning account for hidden layer representations rather than the output distribution. § 3 makes the analogous claim for outputs.

The closest account to ours applies the structure function to language models. Pan et al. (2025) read LLM training as two-part coding. Their predictions follow from a specific generative model; ours follow from the budgeted optimum for general targets (§ 3). Like us, they trace optima across budgets for a given source and read scaling behavior off such a trace, but they do so just for their generative model. We go beyond this work by deriving per-task behavior from the budgeted optimum of an arbitrary target and placing trained networks on the optimal frontier.

Emergence. Analyses of emergence use models built on composed skills (Arora & Goyal, 2023), quanta learned in order of frequency (Michaud et al., 2023), and independently learned skills (Nam et al., 2024). Others model stagewise transitions within training through modes learned in order of strength (Saxe et al., 2014), induction heads (Olsson et al., 2022), and grokking (Varma et al., 2023). Each account takes the skill structure, and often the order of acquisition, as given. Varma et al. (2023) and Wurgaft et al. (2025) come closest, trading loss against circuit efficiency and against the complexity of a fixed set of candidate strategies, respectively. Our Proposition 1 maps these stages onto optima at smaller budgets, and our Prediction 2 then orders acquisitions by loss removed per nat, which reduces to the frequency order of Michaud et al. (2023) when that ratio is equal across tasks. It also makes each acquisition abrupt, because a new prototype pays its full description cost.

Language models acquire finite knowledge, at most about 2 bits per parameter (Allen-Zhu & Li, 2025), and a capacity-limited network retains features in order of importance and sparsity (Elhage

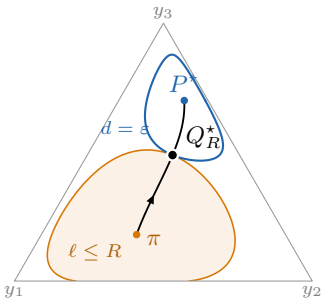

Figure 1: Resource trade-offs illustrated in the simplex of distributions over three outcomes, y_1, y_2, y_3, which holds the target $P^\star$ and the default π; the plot writes d for $\mathrm{KL}(P^\star\|Q)$, ℓ for $\mathrm{KL}(Q\|\pi)$, and $Q^\star_R$ for $Q_{0.6}$. (Blue) Distributions $\mathrm{KL}(P^\star\|Q) = \varepsilon = 0.11$ from the target. (Orange) Distributions describable in $\mathrm{KL}(Q\|\pi) \leq R = 0.28$ from the default. (Black) Admissible models $\{Q_\lambda : \lambda \in [0, \infty)\}$, where $Q_\lambda = \arg\min_Q \left[\mathrm{KL}(P^\star\|Q) + \lambda\,\mathrm{KL}(Q\|\pi)\right]$ and $Q_{0.6}$ (black point).

et al., 2022). In terms of our model, capacity bounds the budget, and importance is the loss removed per task. A fact's accuracy rises with its frequency in pretraining (Kandpal et al., 2023; Mallen et al., 2023). Models learn facts in phases, with a plateau before they acquire precise facts, and hallucinations appear along with knowledge (Zucchet et al., 2025). Several results constrain hallucination (Kalai & Vempala, 2024; Kang et al., 2024). We analyze this behavior as Prediction 3.

3 Modeling Behavioral Scaling Laws

Consider a target distribution, $P^\star$, and an approximate model Q on the same support, with loss $d(Q) = \mathrm{KL}(P^\star\|Q)$. For any loss, $d(Q) = \varepsilon > 0$, there exists an equivalence class of satisfactory distributions, as loss underdetermines behavior (Figure 1, blue). We model resources as an informational bound (Ortega & Braun, 2013): a budget of R nats on $\ell(Q) = \mathrm{KL}(Q\|\pi)$, the nats of description above an easily attainable default, $\pi : \ell(\pi) = 0$, affords a restricted set of models, $S_R = \{Q : \ell(Q) \leq R\}$ (Figure 1, orange). For most $P^\star$ of interest, the budget excludes the desired distribution, $\ell(P^\star) > R$. In the probabilities $Q(y)$ (the coordinates of Figure 1), however, both d and ℓ are convex (Cover & Thomas, 2006). So for any R there exists a unique optimum, $Q^\star_R = \arg\min_{Q \in S_R} d(Q)$. Write $\varepsilon(R) = d(Q^\star_R)$. At $Q^\star_R$, the level set of the loss, $d = \varepsilon(R)$, touches the budget set S_R (Figure 1, black point). The loss this affords, $\varepsilon(R)$, is the Pareto frontier. As R grows, $Q^\star_R$ moves from π to $P^\star$ along the path of admissible models (Figure 1, black curve). $Q^\star_R$ also minimizes the following tempered objective (Boyd & Vandenberghe, 2004):

$$Q^\star_R = \arg\min_Q \left[d(Q) + \lambda\,\ell(Q)\right], \qquad \lambda = -\varepsilon'(R). \tag{1}$$

It is zero when $R \geq \ell(P^\star)$, yielding $P^\star$, and grows without bound as $R \to 0$, yielding π, and falls as R grows. We write $Q^\star(P; \lambda)$ for the minimizer of Eq. (1) with target P in place of $P^\star$ (App. A). Under the model, a learner with R nats minimizes loss within its budget, finding $Q^\star_R$. Proposition 1 (App. B) makes this precise for trained networks: if the outputs a network can reach form the budget set S_R, every local minimum of its loss lies at $Q^\star_R$.

In natural language, constituents (tasks and task families) are manifold and heterogeneous (Radford et al., 2019; Xie et al., 2022; Michaud et al., 2023; Raventós et al., 2023), constituent frequencies are Zipfian (Zipf, 1949; Piantadosi, 2014; Chan et al., 2022), and tasks nest into hierarchies (Saxe et al., 2019). In this setting resources control which tasks are differentiated, when, and to what extent. For hierarchical processes: a family $f \sim \rho_F$, a task $t \mid f \sim \rho(\cdot \mid f)$, and an outcome $y \mid t \sim P^\star_t$, in which task t has prior probability $\rho(t) = \sum_f \rho_F(f)\,\rho(t \mid f)$.

We model the shared structure of tasks via *prototypes* that capture subdistributions. A prototype distribution q_k carries a mass m_k (c.f. Saeedi et al., 2017). A router $p(k \mid t)$ sends tasks to prototypes, and the model serves task t with $Q_t = \sum_k p(k \mid t)\,q_k$. The masses are the router's marginal, $m_k = \sum_t \rho(t)\,p(k \mid t)$. We call the tasks that a prototype serves its *cell*. With unbounded resources, the family reproduces any target, and so the model is minimally constraining. Bounded resources control how many and which prototypes carry mass. The loss charges each query the divergence of its prototype:

$$d(Q) = \sum_t \rho(t) \sum_k p(k \mid t)\,\mathrm{KL}(P^\star_t\|q_k). \tag{2}$$

By convexity, d bounds the mixture loss $\sum_t \rho(t)\,\mathrm{KL}(P_t^\star \| Q_t)$ from above, with equality under deterministic routing. The resource cost has two terms:

$$\ell(Q) = \sum_t \rho(t)\,\mathrm{KL}\big(p(\cdot \mid t) \| m\big) + \sum_k \mathrm{KL}(q_k \| \pi). \tag{3}$$

With one task and one prototype, routing is free and the problem reduces to Eq. (1); without the description term, it becomes the information bottleneck (Tishby et al., 1999; App. C). In the general case the problem is not convex: d couples router and prototypes multiplicatively and ℓ charges each prototype in full. We solve this general version of the problem via iterative updates

$$p(k \mid t) \propto m_k\, e^{-\mathrm{KL}(P_t^\star \| q_k)/\lambda}, \qquad m_k = \sum_t \rho(t)\, p(k \mid t), \qquad q_k = Q^\star\big(\bar{P}_k;\, \lambda/m_k\big), \tag{4}$$

where $\bar{P}_k = \sum_t p(t \mid k)\, P_t^\star$ is the pooled target of prototype k's tasks. Each update minimizes the objective over its block, so the alternation descends monotonically (Neal & Hinton, 1998; App. C). The router update is the Gibbs form, $\arg\min_p \big[-\mathbb{E}_p v + \lambda\,\mathrm{KL}(p\|m)\big] \propto m\, e^{v/\lambda}$ (Donsker & Varadhan, 1975). Here the value is $v = -\mathrm{KL}(P_t^\star \| q_k)$ and the default is m (Blahut, 1972; Tishby et al., 1999). q_k enters the objective only through $m_k\,\mathrm{KL}(\bar{P}_k \| q_k) + \lambda\,\mathrm{KL}(q_k \| \pi)$ (App. C). Each prototype is thus Eq. (1)'s model of the tasks it serves.

3.1 Predictions

In the model, each property of the optimum predicts the behavior of a network at the same rate. The key predictions are as follows, with detailed statements and proofs in the corresponding appendices:

Prediction 1. Power-law scaling (App. D). Assume $\mathcal{D}_n$ is the least loss achievable with n prototypes. If every prototype other than π costs about c nats of description, R nats afford about R/c prototypes, and $\varepsilon(R)$ follows $\mathcal{D}_n$ at $n \approx R/c$. When $\mathcal{D}_n$ decays as $n^{-\gamma}$, $\varepsilon(R)$ decays as $R^{-\gamma}$ up to slowly varying factors. If tasks are Zipfian, $\rho(t) \propto t^{-z}$ with $z > 1$. With equal losses, $\Delta_t = \Delta$, $\mathcal{D}_n = \Delta \sum_{t>n} \rho(t) \propto n^{-(z-1)}$, so $\varepsilon(R) \propto R^{-(z-1)}$ (c.f. Michaud et al., 2023; Hutter, 2021).

Prediction 2. Prediction 2: Metric-dependent emergence (App. E). A cell keeps one prototype until λ falls below a threshold set by how much its tasks reveal about outcomes. Coarse distinctions therefore precede fine ones when families differ more than their constituents do (c.f. Saxe et al., 2019). Describing a new prototype costs a fixed amount, so with each split new prototypes appears at a finite distance from their ancestors. Prototypes, and with them model behavior, thus change discontinuously. Deterministic annealing charges nothing per codevector, and at a critical temperature it splits a cluster by bifurcating along its principal axis (Rose, 1998). The description cost turns this into a jump while aggregate loss remains smooth (Schaeffer et al., 2023, c.f.). When tasks are atomic, and acquiring task t costs c_t nats, task t is allocated a prototype once λ falls below $\lambda_t = \rho(t)\Delta_t/c_t$, the per nat loss recovered. Tasks enter in decreasing order of λ_t. Each entry drops that task's loss by Δ_t (c.f. Michaud et al., 2023).

Prediction 3. Lawful hallucination (App. F) For any target, a task t° without its own prototype is routed to $Q_R^\star(\cdot \mid t^\circ) = \sum_k p(k \mid t^\circ)\, q_k$ (Eq. 4). Let $g_k = \mathrm{KL}(P_{t^\circ}^\star \| q_k) - \lambda \log m_k$. The surrogate concentrates on $k^\circ = \arg\min_k g_k$ once every other prototype's gap $g_k - g_{k^\circ}$ is large next to λ . Membership thus weighs cell size m_k. Each prototype is $Q^\star(\bar{P}_k; \lambda/m_k)$, its pooled target. This has four consequences: *Content.* A task outside the budget is routed to the prototypical answer of its cell. *Frequency.* A cell's prototype sits on the path from Eq. (1) at temperature λ/m_k, so rarer cells answer closer to the default (App. F; c.f. Kandpal et al., 2023; Mallen et al., 2023). *Resolution.* A cell keeps one prototype until λ falls below its threshold (Prediction 2), so cells whose tasks reveal little about outcomes keep a shared surrogate longest. *Rate.* For atomic Zipfian tasks (Prediction 1), surrogates serve task mass $\varepsilon(R)/\Delta \propto R^{-(z-1)}$, so hallucination falls with the same exponent as loss (c.f. Kalai & Vempala, 2024).

4 Experiment 1: Compressing a Library of Markov Tasks

To quantify the above predictions exactly, the target's optimum must be computable. Experiment 1 uses Markov tasks on a Zipfian tree to test § 3 and Predictions 1 and 2. A task is an order-5 Markov

chain over three symbols. Its context is the last five symbols, so a task is a table of next-symbol distributions over $C = 3^5 = 243$ contexts. The contexts form 81 blocks of three that share their four most recent symbols. Every block is doubly stochastic: its entries sum to 1 over the next symbol and over the oldest symbol, so (1) every context is equally frequent under every task and (2) given its four most recent symbols or fewer, the next symbol is uniform. A document is 64 symbols from one task, so 59 transitions have a full context, too few to estimate an unseen task in context.

The tasks form a tree. A child at depth h moves $0.2 \times 0.9^{h-1}$ nats per token from its parent, along a random direction. Each node keeps 31% of its need and passes 69%, $3^{-1/3}$, to its children. Branches stop below need 2×10^{-5}. Depth h then holds about 3^h tasks of need about $3^{-4h/3}$, so need falls as rank$^{-4/3}$ (fitted slope -1.39; Prediction 1). Content, $\ell(P_\theta)/C$ nats per token, rises with depth: common tasks are near uniform, and rare ones are peaked. We hold out 64 leaves and 4 subtrees (216 tasks: the 64 leaves and 152 in the subtrees), which leaves 2,372 trained tasks, and include the following control task libraries. *Flat* redraws every table independently at the same need and content, so tasks share no structure. *Order 4* has 81 contexts, so in-context estimation is possible. *Twins* repeat the recipe with fresh random draws.

On this library, Eqs. (2)–(3) read task θ through Q_j at $d(\theta, j) = \frac{1}{C}\sum_c \mathrm{KL}(P_\theta(\cdot|c)\|Q_j(\cdot|c))$ nats per token. It charges each prototype $\ell(Q_j) = \sum_c \mathrm{KL}(Q_j(\cdot|c)\|u)$ nats, with u the uniform default. We sweep λ over 64 values from 10^{-11} to 10. The solver starts from one cell per task and alternates assignment with the prototype update of Eq. (4). It deletes any cell whose rent $\lambda\ell(Q_j)$ exceeds what its members would lose at their next-best cells. Once at most 512 cells remain, it also merges any pair of cells whose union lowers the tempered objective. The solver routes deterministically and omits the routing term, which adds at most $\log K$ nats. The sweep traces the frontier $\varepsilon(R)$.

The Bayesian reader with access to all tasks, with the needs as its prior, sets the floor to $L^\star$, the entropy rate plus the cost of identifying the task, averaged over positions 32–63, where every reader has seen at least 27 transitions. § 3 puts a model with R nats on the frontier:

$$L(N) = L^\star + \varepsilon\big(R(N)\big). \tag{5}$$

This charges every network the ideal reader's cost of identification. Suppose the frontier falls as $R^{-(z-1)}$, as in Prediction 1's atomic case, and the rate a network reaches grows with its size as $R(N) \propto N^a$. Then $L - L^\star \propto N^{-a(z-1)}$.

4.1 Results

We train decoder-only transformers (Vaswani et al., 2017; Radford et al., 2019) across a range of parameter counts and hyperparameters (App. G). To project the models onto the information plane, we run the network on 64 probe documents per task. Averaging its posteriors over positions 32–63 per (task, context) pair yields each task's implied table for that model. The solver groups the implied tables across temperatures. Each grouping speaks through its need-weighted cell means, and we measure those prototypes against the true tables. The network's rate R is that of the cheapest grouping whose distortion lies within 2% of that of the ungrouped implied tables.

Rate grows with size, as $R \propto N^{0.82}$ across 3.5 decades, and it rises during training (Figure 2A). Eq. (5) predicts each network's loss from its rate (Figure 2B). $L - L^\star$ falls from 0.18 to 0.032 nats per token. $\varepsilon(R)$ at each network's rate follows it. $\varepsilon(R)$ falls with slope -0.28, close to the tie $-a(z-1) = -0.82/3 = -0.27$, and $L - L^\star$ with slope -0.23. Networks sit on the frontier, as § 3 predicts (Figure 2D). The frontier's smoothness hides step-like acquisitions. Each task's excess stays near its content until the optimum gives it a prototype, then drops, heavy tasks first, as Prediction 2 predicts. At a given rate, networks group tasks as the optimum does. The adjusted mutual information between a network's cells and the optimum's (Vinh et al., 2010) is 0 at chance and sits at 0.73–0.85 above 10^4 parameters (Figure 2C). The optimum with the network's cell count agrees as well. Below 10^4 parameters, networks hold one or two cells, and the index degenerates. On the simplex, learned cells and prototypes resemble the optimum's (Figure 3).

As networks grow, they acquire tasks in the optimal order. A task's acquired fraction is $h_\theta = 1 - \sqrt{d_\theta\, C/\ell(P_\theta)}$, clipped to $[0, 1]$. For the optimum, d_θ is the misfit of the task's prototype; for a network, it is the task's window excess. Because KL grows quadratically with distance, h_θ is the share of the way from the uniform table to the task's own table that its reading covers (Figure 2E).

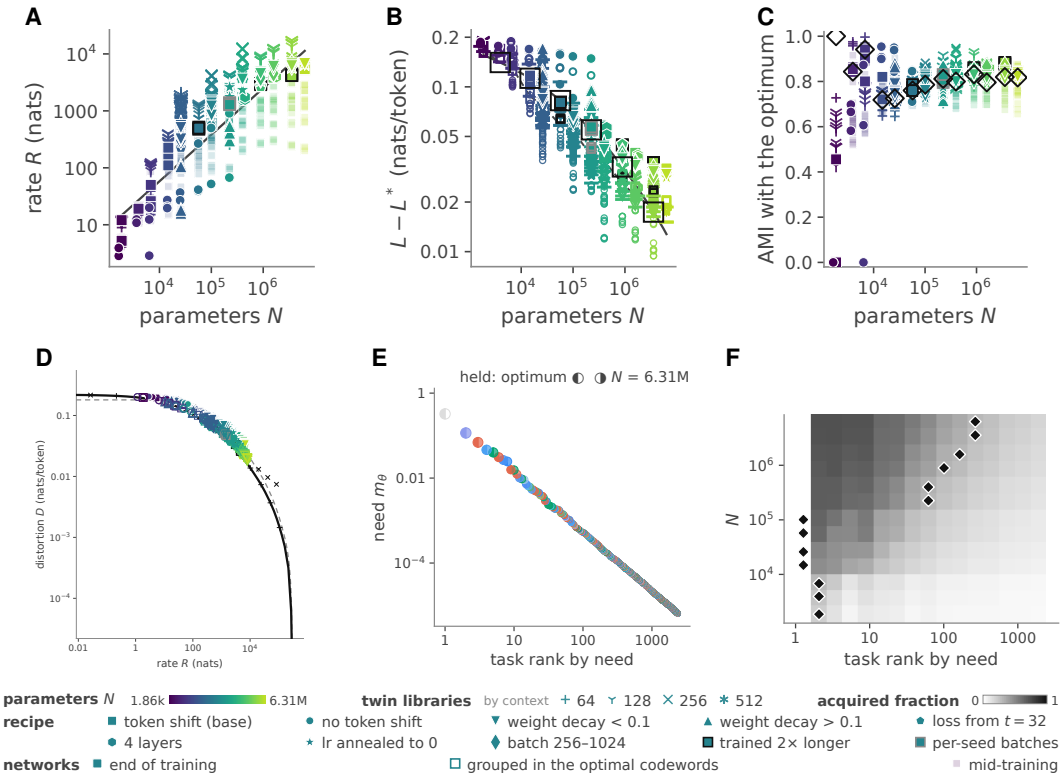

Figure 2: Transformers trained on Markov tasks. **(A) Rate increases with size.** Rate by size. The line is a power law fitted on alternate sizes, and faint markers are earlier checkpoints. **(B) The optimum at a network's rate predicts its loss.** $L - L^\star$ (filled) by size, beside $\varepsilon(R)$ at each network's measured rate (small open). The line is $\varepsilon\big(R(N)\big)$, with $R(N)$ the power law fitted in panel A. Large open markers are predictions at the sizes left out of the fit. Dashes are the optimum's Bayes reader at each network's rate, identification included. **(C) Networks group tasks optimally.** Adjusted mutual information between each network's cells and the optimum's at the network's rate (1: identical groupings; 0: chance). Diamonds: the optimum with the network's cell count. **(D) Networks approach the frontier.** The thick curve is the frontier $\varepsilon(R)$. The dashed curve is a tree-aware baseline that reads each task through its deepest acquired ancestor. Networks (filled) and their groupings in the optimal prototypes (open) approach the frontier. Ink dots are the optima of Figure 3. $+$ marks a known codebook, and $\times$ marks where the readout places it. **(E) Networks acquire tasks in the optimal order.** Needs by rank. Each dot's left half fills with how much the optimum at the largest network's rate acquires the task. Its right half fills with how much the 6.31M network does. **(F) Lawful emergence.** Acquired fraction by size (rows) and need-rank bin (columns), white for not acquired and black for acquired. Diamonds mark where the optimum at each size's rate falls below half acquired.

Across sizes, the acquired region extends to rarer tasks as N grows. At large sizes it tracks the rank where the optimum at each size's rate falls below half (Figure 2F).

5 Experiment 2: Factual Recall in Pretrained Language Models

At pretraining scale the optimum is intractable, so we test the theory on a component of the language-modeling loss with an approximately solvable optimum: factual recall (c.f. Petroni et al., 2019). In terms of Section 3, a factual question is a task whose target is a point mass on its answer. A model either acquires the fact or reads it through a prototype. The theory makes three predictions: A model acquires a fact when its need exceeds a temperature that falls with the model's size (Prediction 2; Figure 5A, B). A model that does not acquire a fact answers with its cell's typical answer (Prediction 3; Figure 6). A model acquires broad cells before their constituents (Prediction 2; Figure 6).

5.1 Setup

Each question has one true answer and belongs to one of four families: *Researcher* $\rightarrow$ *area*: "In which subfield of computer science is X known for their research?", answered with one of the 27

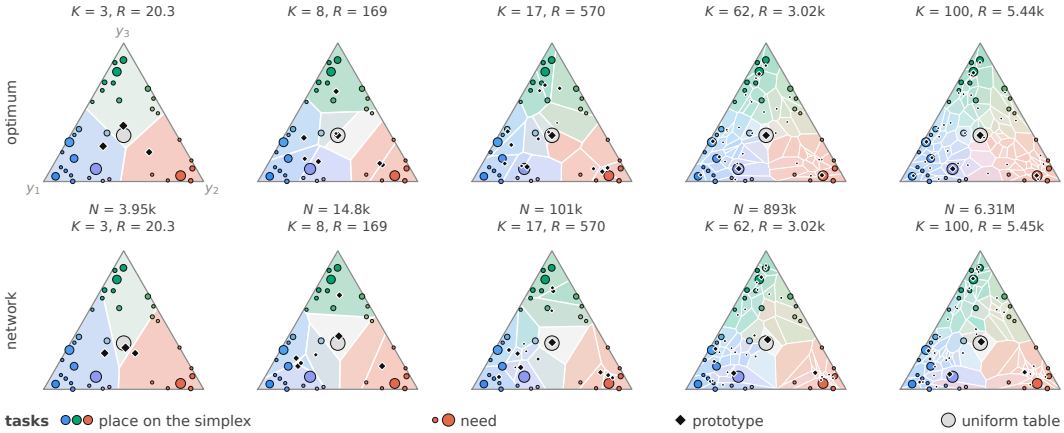

Figure 3: **Networks compress optimally.** The panel uses the context where the task tables spread widest. Tasks are located on the simplex by their next-symbol distributions, with area growing with need and a color it keeps throughout. Diamonds are prototypes, shaded regions are the tasks each prototype claims. Top: the optimum with the network's cell count and rate. Bottom: the networks' learned compressions.

areas of CSRankings (Berger, 2025). The cell is "a computer scientist". *Researcher* → *institution*: "At which university does the computer scientist X work?", for the same researchers and cell. *Item* → *founding year*: universities, museums, football clubs, journals and bridges. The cell is the class and country, or the class alone for journals. An answer counts if it lies within a year of an accepted date. *Graph relations*: a country's currency and official language, a river's mouth, a film's director, and a mountain's country. The cell is the continent, the river's countries, or the film's decade.

A question's need w_t is its subject's citation count (OpenAlex; Priem et al., 2022) or number of Wikipedia sitelinks (Wikidata; Vrandečić & Krötzsch, 2014), plus one, times a factor u per property.[1] $p_0(t)$ is the probability that a receiver who knows only the question's cell puts on the truth. We estimate it leave-one-out from the cell's other constituents. For founding years, a small cell borrows from its class. The cell's law is its prototype, the pooled target $\bar{P}_k$ of Eq. (C.2). $\ell_t = -\ln p_0(t)$. We draw subjects with equal quotas over ten bins of log need: 400 researchers, and 100 subjects per class and per relation, for up to 1,800 questions.

A question's target $P_t^\star$ is a point mass on its answer, so $\mathrm{KL}(P_t^\star \| \bar{P}_k) = \ell_t$ sets both sides of Prediction 2's conversion. It is the cost of describing the fact above its cell's prototype, and also the loss each use pays when the model reads the fact through that prototype. With loss weighted by need w_t,

[1] 1.0 for areas, 0.35 for institutions, 0.3 for universities, 0.1 for museums, 0.2 for football clubs, journals and bridges, 0.8 for currencies and languages, 0.05 for river mouths, 0.3 for film directors, and 0.5 for mountains' countries.

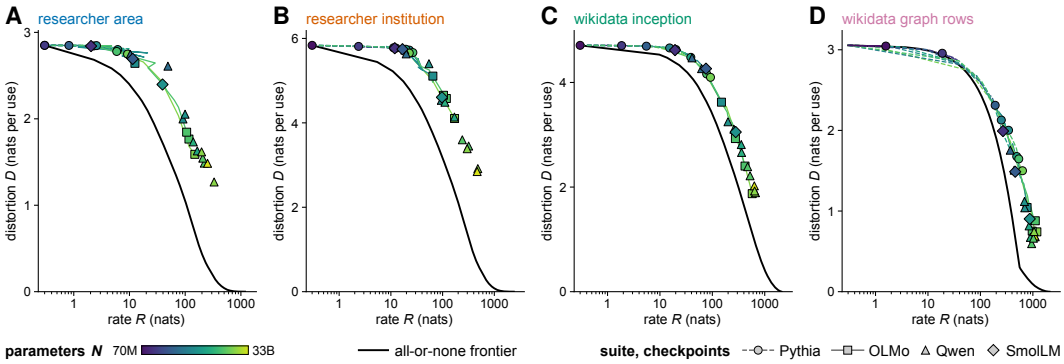

Figure 4: Pretrained language models on the rate–distortion plane of factual recall. **(A–D) Models approach the frontier.** Rate R counts the nats a model acquires, and distortion D the need-weighted loss per use. The black curve is the frontier. Markers are the 26 models, colored by size from purple (small) to yellow (large) and shaped by suite: circles for Pythia, squares for OLMo, triangles for Qwen, diamonds for SmolLM. Lines trace training checkpoints, dashed for Pythia and solid for OLMo. Rates at or below 0.3 nats sit on the left edge.

Prediction 2's critical temperature reduces to the fact's need:

$$\lambda_t = \frac{w_t\,\ell_t}{\ell_t} = w_t. \tag{6}$$

The temperature is thus the least need at which an acquired nat pays for itself. At temperature λ, the optimum acquires the facts with $w_t \geq \lambda$, in order of need and whatever their content. This neglects routing, which Prediction 2 bounds and which costs little for rare facts. A receiver that acquires fact t with probability h_t spends $h_t\ell_t$ nats on it and misses $(1-h_t)\ell_t$ nats per use. Acquiring the k most needed facts gives the vertex $R_k = \sum_{i\leq k}\ell_i, D_k = \sum_{i>k}\rho(i)\,\ell_i$, where $\rho(t) = w_t/\sum_{t'} w_{t'}$ is the fact's share of its family's need. Acquiring the next fact with probability between zero and one moves along the segment to the next vertex. The popularity effects of factual recall (Kandpal et al., 2023; Mallen et al., 2023) follow, with the threshold at λ.

5.2 Models

We analyze 26 open base models: Pythia, deduplicated, at eight sizes from 70M to 12B (Biderman et al., 2023); OLMo 2 at 1B, 7B and 13B (Team OLMo et al., 2024); Olmo 3 at 7B (Team Olmo et al., 2025); Qwen2.5 from 0.5B to 32B (Qwen Team et al., 2024); Qwen3 from 0.6B to 14B (Yang et al., 2025); SmolLM2 from 135M to 1.7B (Ben Allal et al., 2025). Together they span 70M to 33B parameters and 0.3T to 36T training tokens. We also read 97 training checkpoints: eight per Pythia size, and 33 from OLMo 2's first pretraining stage. We place five fixed question–answer pairs before each question (Brown et al., 2020). We then read the probability $Q(y_t)$ that the model's reply begins with a correct surface form.[2] For founding years, $Q(y_t)$ is the mass on years within a year of the truth. A receiver that acquires a fact with probability h, and otherwise reads the cell's prototype, has mass $p_0 + (1-p_0)h$ on the truth. A model's acquired fraction inverts this: $h_t = \frac{Q(y_t)-p_0(t)}{1-p_0(t)}$, clipped to $[0,1]$. The model then occupies its family's information plane at $R = \sum_t h_t\ell_t$ and $D = \sum_t \rho(t)(1-h_t)\ell_t$.

One forecast covers every (model, question) cell:

$$\hat{Q}(y_t) = p_0 + (1-p_0)\,\sigma\!\left(\frac{\ln \kappa w_t - \ln \lambda(N)}{s}\right), \qquad \lambda(N) = \bar{\lambda}\left(\frac{N}{10^9}\right)^{g}. \tag{7}$$

It is the optimum's step at $\kappa w_t = \lambda(N)$ (Eq. 6), softened by width, s. Experiment 1 fitted rate as a power law in size; here we fit temperature as a power law in size. We fit four numbers by cross-entropy over the final models' cells: $\bar{\lambda} = 1.25 \times 10^6$, the temperature at 10^9 parameters, in citation-equivalents per acquired nat; $g = -1.09$, how λ falls with size; $\kappa = 4.2 \times 10^3$, citations per sitelink, which puts both need proxies on one scale ($\kappa = 1$ on citation rows); $s = 1.50$, the width.

A question's acquisition size N_{50} is where a logistic in $\ln N$, fit to the final models' acquired fractions, crosses one half. The fitted forecast, Eq. (7), predicts $N_{50} = \lambda^{-1}(\kappa w_t)$. For a founding-year question, a context score places the model's wrong years between the class-wide prototype and the cell's. A cell's N_{50} is where its constituents' mean score crosses half. Taking a cell's need as the sum of its constituents', the fitted forecast predicts that size too.

5.3 Results

In all four families, models close a median 29–54% of the gap between a need-blind store and the frontier at their rate (areas 36%, institutions 29%, founding years 53%, graph relations 54%; Figure 4). Models at one rate sit at the same distortion across families and training stages.

Extent of acquisition follows need/λ (Figure 5A). We fit each model's λ on each family, with the width fixed at the jointly fitted s, and divide each question's need by it. Acquisition extent then collapses onto a single curve. The fit centers each family on $w = \lambda$. Below the temperature the families agree: a question used a tenth as often as the temperature has an acquired fraction of 0.14–0.22 in every family, lowest for founding years. The rise is gradual, with one width across families, and at ten times λ the acquired fraction reaches 0.75 (graph relations only; the logistic gives 0.82). Split into content terciles, acquired fraction at $w = \lambda$ is 0.51, 0.60 and 0.70, so content does not cancel fully (Eq. 6).

[2] We choose the forms so that no form's tokens begin another's and no form names another answer, and we sum their probabilities (Holtzman et al., 2021).

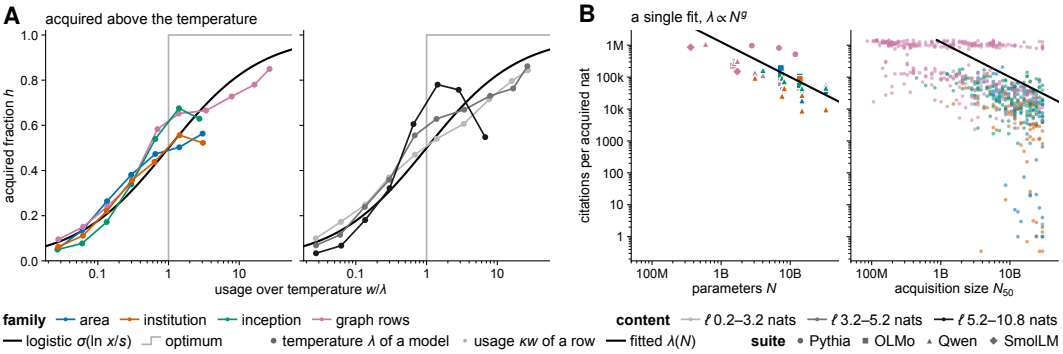

Figure 5: **(A) Acquired fraction follows need over temperature.** We fit each model's λ on each family with the width fixed at the jointly fitted s of Eq. (7), omit temperatures outside the family's range of need, divide every question's need by that temperature, and bin over models and questions. The black curve is the logistic with that width; the grey step is the optimum. *Left:* by family. *Right:* by content: cheap (0.2–3.2 nats), such as a European country's currency, where the euro is a good guess; medium (3.2–5.2 nats); and expensive (5.2–10.8 nats). **(B) Temperature falls with size.** *Left:* each final model's λ on each family (blue: areas; orange: institutions; green: founding years; pink: graph relations), in citation-equivalents per acquired nat, by parameter count, shaped by suite as in Figure 4, with the fitted $\lambda(N)$ of Eq. (7). *Right:* each question's need by its acquisition size N_{50}, the size at which the final models have half acquired it, for questions acquired within the range of sizes, with the acquisition size Eq. (7) predicts, $\lambda^{-1}(\kappa w)$.

Temperature falls with size (Figure 5B). Each final model's λ falls with size as the power law $\lambda \propto N^g$. The acquisition sizes, $\lambda^{-1}(\kappa w_t)$, track the observed N_{50} ($r = 0.59$ on log scales, $n = 660$; median error a factor of 5). Acquisition runs coarse to fine (Figure 6). Past its cell's N_{50}, a model that has not yet acquired a founding year answers with its cell's typical years (Prediction 3). A cell's need is the sum of its constituents', so cells are predicted to be acquired before their constituents; the cell comes first in 66% (27 of 41) of observed pairs, and in 90% (123 of 137) if constituents that never cross count as later. In Figure 6, 16 of the 17 cells that cross one half within the range of sizes do so before their median constituent, a constituent that never crosses counting as later; the exception is museums in the UK.

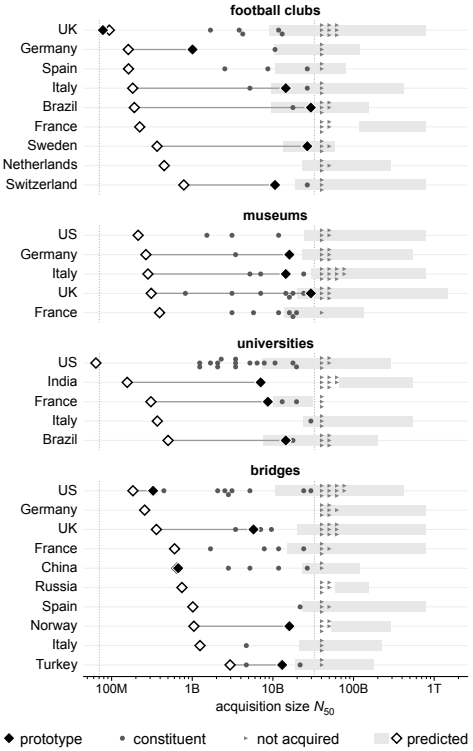

Figure 6: **Cells are acquired before their constituents.** Each row is a cell (class and country, four or more questions): the size at which it is half acquired (filled diamond; none if outside the range of sizes, dotted) and the sizes at which its constituents' founding years are half acquired (dots; triangles if never). The hollow diamond is the acquisition size Eq. (7) predicts from the cell's need, the sum of its constituents'; the pale band is the range it predicts from the constituents' own needs.

6 Conclusion

We model a language model as the optimum of an objective set jointly by the data and a budget. This path predicts power-law loss (Prediction 1), abrupt acquisition ordered by loss removed per nat (Prediction 2), and, for tasks outside the budget, answers drawn from their cell's prototype (Prediction 3). In Experiment 1, transformers approach the frontier at their rate, group tasks as the optimum at that rate does, and acquire tasks in its order. In Experiment 2, pretrained models close the gap between a need-blind store and the frontier at their rate. There, a temperature falling as a power law in size forecasts which facts each model acquires, and models acquire cells before their constituents. Behavior that loss leaves underdetermined thus becomes, in part, predictable from the rate and the optimum it affords. However, our analysis remains limited in the set of models evaluated, which we hope to build on in future research.

7 AI USE STATEMENT

Generative AI tools assisted with parts of this work. Their use was limited to improving the clarity and grammar of the manuscript and identifying related literature. The authors checked every AI-assisted contribution: edited text was reviewed for accuracy, AI-assisted code was tested by at least one author, and all references were verified against their original sources. The authors take full responsibility for all content in this paper.

REFERENCES

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. *Journal of Machine Learning Research*, 25(331):1–58, 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FxNNiUgtfa.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society and Oxford University Press, Providence, RI, 2000. doi: 10.1090/mmono/191.

Venkat Anantharam, Amin Gohari, Sudeep Kamath, and Chandra Nair. On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover. arXiv:1304.6133, 2013. A conference version appeared as "On hypercontractivity and a data processing inequality" in *2014 IEEE International Symposium on Information Theory*, pp. 3022–3026, 2014.

John R. Anderson. The adaptive nature of human categorization. *Psychological Review*, 98(3): 409–429, 1991. doi: 10.1037/0033-295X.98.3.409.

Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. arXiv:2307.15936, 2023.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big — Data-centric training of a fully open small language model. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=3JiCl2A14H. arXiv:2502.02737.

Emery D. Berger. CSRankings: Computer science rankings. https://csrankings.org, 2025.

Dimitri P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 2 edition, 1999.

Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. Establishing task scaling laws via compute-efficient model ladders. In *Conference on Language Modeling*, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin 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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0561738a239a995c8cd2ef0e50cfa4fd-Abstract-Conference.html.

Nicholas H. Bingham, Charles M. Goldie, and Jozef L. Teugels. *Regular variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, UK, 1987. doi: 10.1017/CBO9780511721434.

Richard E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972. doi: 10.1109/TIT.1972.1054855.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004. doi: 10.1017/CBO9780511804441.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/77c6ccacfd9962e2307fc64680fc5ace-Abstract-Conference.html.

Tyler A. Chang and Benjamin K. Bergen. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16, 2022. doi: 10.1162/tacl_a_00444.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8281–8297. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.568.

Henry C. Conklin, Tom Hosking, Yi-Chern Tan, Jonathan D. Cohen, Sarah-Jane Leslie, Thomas L. Griffiths, Max Bartolo, and Seraphina Goldfarb-Tarrant. Learning is forgetting; LLM training as lossy compression. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tvDlQj0GZB. Extended version: arXiv:2604.07569.

Robert M. Corless and David J. Jeffrey. The Wright ω function. In Jacques Calmet, Belaid Benhamou, Olga Caprotti, Laurent Henocque, and Volker Sorge (eds.), *Artificial Intelligence, Automated Reasoning, and Symbolic Computation*, volume 2385 of *Lecture Notes in Computer Science*, pp. 76–89, Berlin, Heidelberg, 2002. Springer. doi: 10.1007/3-540-45470-5_10.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, NJ, 2 edition, 2006. doi: 10.1002/047174882X.

Imre Csiszár and Gábor Tusnády. Information geometry and alternating minimization procedures. *Statistics & Decisions*, Supplement Issue 1:205–237, 1984.

Łukasz Dębowski. From Zipf's law to neural scaling through Heaps' law and Hilberg's hypothesis. *Journal of Machine Learning Research*, 27(194):1–35, 2026.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jznbgiynus.

Monroe D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975. doi: 10.1002/cpa.3160280102.

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*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5f1eee2509599faeeb3570a887016a64-Abstract-Conference.html.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, Eran Malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning Markov chains. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/75b0edb869e2cd509d64d0e8ff446bc1-Abstract-Conference.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html. arXiv:2209.10652.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Luca Soldaini, Jenia Jitsev, Alex Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, Thomas Kollar, et al. Language models scale reliably with over-training and on downstream tasks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iZeQBqJamf.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernion, Shauna Kravec, Ben Mann, et al. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pp. 1747–1764. ACM, 2022. doi: 10.1145/3531146.3533229.

Micah Goldblum, Marc Finzi, Keefer Rowan, and Andrew Gordon Wilson. Position: The no free lunch theorem, Kolmogorov complexity, and the role of inductive biases in machine learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15788–15808. PMLR, 2024.

Siegfried Graf and Harald Luschgy. *Foundations of quantization for probability distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 2000. doi: 10.1007/BFb0103945.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. arXiv:1712.00409, 2017.

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pp. 5–13. ACM, 1993. doi: 10.1145/168304.168306.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html. Preprint: "Training Compute-Optimal Large Language Models", arXiv:2203.15556.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.564.

Marcus Hutter. Learning curve theory. arXiv:2102.04074, 2021.

Hong Jun Jeon and Benjamin Van Roy. Information-theoretic foundations for neural scaling laws. arXiv:2407.01456, 2024.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 160–171. ACM, 2024. doi: 10.1145/3618260.3649777.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 2023.

Katie Kang, Amrith Setlur, Claire Tomlin, and Sergey Levine. Deep neural networks tend to extrapolate predictably. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ljwoQ3cvQh.

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:2001.08361, 2020.

Tamás Linder, Ram Zamir, and Kenneth Zeger. High-resolution source coding for non-difference distortion measures: Multidimensional companding. *IEEE Transactions on Information Theory*, 45(2):548–561, 1999. doi: 10.1109/18.749002.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22188–22214. PMLR, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Sanae Lotfi, Marc Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32801–32818. PMLR, 2024.

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)*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546.

Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, Alisa Liu, et al. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=DwgRm72GQF.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5b6346a05a537d4cdb2f50323452a9fe-Abstract-Conference.html.

Yoonsoo Nam, Nayara Fonseca, Seok Hyeong Lee, Chris Mingard, and Ard A. Louis. An exactly solvable model for emergence and scaling laws in the multitask sparse parity problem. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/45f7ad60c01f17111ccd8ac2f2fb77e3-Abstract-Conference.html.

Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan (ed.), *Learning in Graphical Models*, volume 89 of *NATO ASI Series, Series D: Behavioural and Social Sciences*, pp. 355–368. Kluwer Academic Publishers, Dordrecht, 1998. doi: 10.1007/978-94-011-5014-9_12.

Maher Nouiehed and Meisam Razaviyayn. Learning deep models: Critical points and local openness. *INFORMS Journal on Optimization*, 4(2):148–173, 2022. doi: 10.1287/ijoo.2021.0062.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html. arXiv:2209.11895.

OpenAI. GPT-4 technical report. arXiv:2303.08774, 2023.

Pedro A. Ortega and Daniel A. Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683, 2013. doi: 10.1098/rspa.2012.0683.

David Owen. How predictable is language model benchmark performance? arXiv:2401.04757, 2024.

Zhixuan Pan, Shaowen Wang, Pengfei Liao, and Jian Li. Understanding LLM behaviors via compression: Data generation, knowledge acquisition and scaling laws. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f61d7778e89b9221d1ea0ce8428b7014-Abstract-Conference.html.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander 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)*, pp. 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-1250.

Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, 2014. doi: 10.3758/s13423-014-0585-6.

Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. In *26th International Conference on Science, Technology and Innovation Indicators (STI 2022)*, Granada, Spain, 2022. doi: 10.5281/zenodo.6936227. arXiv:2205.01833.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, et al. Qwen2.5 technical report. arXiv:2412.15115, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2e10b2c2e1aa4f8083c37dfe269873f8-Abstract-Conference.html.

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.

Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10(3–4):441–451, 1959. doi: 10.1007/BF02024507.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. doi: 10.1016/0005-1098(78)90005-5.

Kenneth Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998. doi: 10.1109/5.726788.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1cded4f97cf5f01a284c574110b7e3b9-Abstract-Conference.html.

Ardavan Saeedi, Tejas D. Kulkarni, Vikash K. Mansinghka, and Samuel J. Gershman. Variational particle approximations. *Journal of Machine Learning Research*, 18(69):1–29, 2017.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://openreview.net/forum?id=_wzZwKpTDF_9C.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019. doi: 10.1073/pnas.1820226116.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/adc98a266f45005c403b8311ca7e8bd7-Abstract-Conference.html.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. Continued in 27(4), 623–656.

Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, 7(4):142–163, 1959. Reprinted in R. E. Machol (ed.), Information and Decision Processes, McGraw-Hill, 1960, 93–126.

Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

Susanne Still, James P. Crutchfield, and Christopher J. Ellison. Optimal causal inference: Estimating stored information and approximating causal architecture. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(3):037111, 2010. doi: 10.1063/1.3489885.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

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, et al. 2 OLMo 2 Furious. arXiv:2501.00656, 2024.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, et al. Olmo 3. arXiv:2512.13961, 2025.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, Monticello, IL, 1999.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. arXiv:2309.02390, 2023.

Aditya Varre, Gizem Yüce, and Nicolas Flammarion. Learning in-context n-grams with transformers: Sub-n-grams are near-stationary points. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60924–60963. PMLR, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Nikolai K. Vereshchagin and Paul M. B. Vitányi. Kolmogorov's structure functions and model selection. *IEEE Transactions on Information Theory*, 50(12):3265–3290, 2004. doi: 10.1109/TIT.2004.838346.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010.

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014. doi: 10.1145/2629489.

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. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. Inverse scaling can become U-shaped. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15580–15591. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.963.

Tailin Wu, Ian Fischer, Isaac L. Chuang, and Max Tegmark. Learnability for the information bottleneck. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 1050–1060. PMLR, 2019. Also published in *Entropy*, 21(10):924, 2019.

Daniel Wurgaft, Ekdeep Singh Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, and Noah D. Goodman. In-context learning strategies emerge rationally. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/abb22b1e4b4ab2988862be9d7ed083f9-Abstract-Conference.html.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13711–13738. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.767.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, et al. Qwen3 technical report. arXiv:2505.09388, 2025.

Oğuz Kaan Yüksel, Rodrigo Alvarez Lucendo, and Nicolas Flammarion. Incremental learning of sparse attention patterns in transformers. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026. URL https://openreview.net/forum?id=vSRh1qU5sH. arXiv:2602.19143.

Paul L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149, 1982. doi: 10.1109/TIT.1982.1056490.

George Kingsley Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, Cambridge, MA, 1949.

Nicolas Zucchet, Jörg Bornschein, Stephanie C. Y. Chan, Andrew K. Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? Dynamics, curricula and hallucinations. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=vBcGnragkr. arXiv:2503.21676.

A THE ONE-QUERY OPTIMUM

Setting: § 3's $P^\star$, π, d, ℓ and Eq. (1). For a fixed value $v(y)$ per outcome, § 3's Gibbs form with default π solves the linear analogue of Eq. (1):

$$\arg\min_{Q} \left[-\mathbb{E}_Q v + \lambda\, \ell(Q)\right] = \pi e^{v/\lambda}/Z \tag{A.1}$$

(Donsker & Varadhan, 1975; Rafailov et al., 2023). The identity holds for any default μ in place of π; App. C and App. F use the router's masses m. The objective is $\lambda\,\mathrm{KL}\big(Q\|\pi e^{v/\lambda}/Z\big) - \lambda \log Z$ with $Z = \sum_y \pi(y) e^{v(y)/\lambda}$, the KL term vanishes at the Gibbs distribution and nowhere else.

The loss d is not linear in Q, but at $Q^\star_R$ Eq. (1) and its version with d linearized at $Q^\star_R$ share a gradient, and the linearized version is convex, so $Q^\star_R$ minimizes it as well. The linearized loss credits each unit of mass on y with the loss recovered, $v(y) = -\partial d/\partial Q(y) = P^\star(y)/Q^\star_R(y)$, so by (A.1)

$$Q^\star_R(y) \;\propto\; \pi(y) \exp \frac{P^\star(y)}{\lambda\, Q^\star_R(y)}. \tag{A.2}$$

Outcomes the optimum underpredicts relative to the target gain the most log-probability over the default.

Write $Q = \pi e^{w-\nu}$, with tilt w and log-normalizer ν. Given ν, Eq. (A.2) asks of each outcome that $w(y) = P^\star(y)/\lambda Q(y)$, that is, $w + \log w = \log \frac{P^\star(y)}{\lambda \pi(y)} + \nu$. The Wright omega function solves this: $w(y) = \omega\big(\log \frac{P^\star(y)}{\lambda\pi(y)} + \nu\big)$ (Corless & Jeffrey, 2002). Then $Q(y) = P^\star(y)/\lambda w(y)$ falls as ν rises, so bisection on the scalar ν finds the unique normalized solution, $Q^\star(P^\star; \lambda)$. Outcomes with $P^\star(y) = 0$ take $w(y) = 0$, the limit $\omega(-\infty) = 0$, so the optimum scales them by one common factor $e^{-\nu}$: its wrong answers keep the default's proportions at every λ. Given w, $\nu = \log \sum_y \pi(y) e^{w(y)}$

renormalizes Q. Each update maximizes $L(w,\nu) = \frac{1}{\lambda}\mathbb{E}_{P^\star} \log w - \nu - \sum_y \pi(y) e^{w(y)-\nu}$ over $w \geq 0$, where excluded outcomes sit at $w = 0$, since the partial derivatives of L vanish at $w = P^\star/\lambda Q$ and at $\sum_y Q(y) = 1$ (Neal & Hinton, 1998; Csiszár & Tusnády, 1984). Because L is jointly concave, neither step can decrease L, and the iterates converge to its maximizer, which satisfies Eq. (A.2) and sums to one (Bertsekas, 1999, Prop. 2.7.1).

Against a uniform default over three outcomes, a point mass costs $\log 3 \approx 1.10$ nats, the information needed to name the outcome; the hedge $(0.2, 0.2, 0.6)$ costs 0.15 nats. In general a point mass δ_y costs $\ell(\delta_y) = -\log \pi(y)$, and answering with π loses the same $-\log \pi(y)$ per query.

Every model satisfies $d(Q) \geq \varepsilon(\ell(Q))$, and a model near the frontier is near the optimum:

$$d(Q) - \varepsilon(R) \;\geq\; \tfrac{1}{2} \min_y P^\star(y)\, \|Q - Q^\star_R\|_2^2 \quad \text{for every } Q \in S_R. \tag{A.3}$$

The bound holds because $Q^\star_R$ satisfies the first-order optimality condition over the convex set S_R, and d is strongly convex on the simplex with modulus $\min_y P^\star(y)$ (Boyd & Vandenberghe, 2004, §§4.2.3, 9.1.2): its Hessian $\mathrm{diag}(P^\star/Q^2)$ dominates $\mathrm{diag}(P^\star)$ because $Q \leq 1$. Fixing R therefore collapses the equal-loss surface of § 3: two models in S_R that each come within δ of $\varepsilon(R)$ lie within $2\sqrt{2\delta/\min_y P^\star(y)}$ of each other.

B Proof of Proposition 1

Setting: one query, as in § 3. Q_θ is the network's output distribution at weights θ, T the tangent space of the simplex, and J the Jacobian of Q_θ in θ.

Proposition 1 (capacity as budget)**.** *Suppose the network reaches the set S_R — the outputs $\{Q_\theta\}$ form S_R — with $\ell(P^\star) > R$, where* $\mathrm{range}\, J = T$ *at interior points of S_R. At boundary points J loses rank, and* $\mathrm{range}\, J$ *is the tangent space of the boundary. Suppose also that weights near any θ reach every output near Q_θ in S_R (local openness; Nouiehed & Razaviyayn, 2022). Then every local minimum of $\theta \mapsto d(Q_\theta)$ lies at $Q^\star_R$, and there the descent direction of d is λ times the outward normal of S_R.*

Gradient descent sees the data gradient through J: the parameter gradient $J^\top \nabla_Q d$ discards the part of $\nabla_Q d$ that points out of the reachable set. The data gradient therefore splits in two: a part along the boundary of S_R, which vanishes at $Q^\star_R$, and a part normal to it. At $Q^\star_R$ its descent direction, λn with n the outward normal, points out of S_R and away from π. Discarding it has the same effect as the resource force $-\lambda n$, pointing toward π, that Eq. (1) applies explicitly. Training stops where the full data gradient points out of the reachable set, so the part the network sees vanishes.

Proof. Let $g = \Pi_T \nabla d$ and $n = \Pi_T \nabla \ell$ be the gradients projected onto T; on the boundary of S_R, n is the outward normal. A local minimum is stationary, $J^\top \nabla d = 0$. At an interior point, $\mathrm{range}\, J = T$ gives $g = 0$ and $Q = P^\star \notin S_R$, so local minima lie on $\ell = R$. There $\mathrm{range}\, J = \{v \in T : \langle n, v\rangle = 0\}$. Stationarity makes g orthogonal to it, so $g = -\lambda n$ for a scalar λ. Nearby weights reach the inward direction $-n$ (local openness), along which d changes at rate $\langle g, -n\rangle = \lambda \|n\|^2$, so a local minimum has $\lambda \geq 0$. The Jacobian conditions alone do not supply this step, since at a boundary point J spans no inward direction. $g + \lambda n = 0$, $\lambda \geq 0$ and $\ell = R$ are the KKT conditions of $\min_{Q \in S_R} d(Q)$, whose stationarity condition is that of Eq. (1). Convexity makes them sufficient (Boyd & Vandenberghe, 2004, §5.5.3), and uniqueness gives $Q = Q^\star_R$. □

Proposition 1 moves the question from optimization to geometry. Capacity acts as the budget of Eq. (1) when the outputs a network can reach form a KL ball around π; for a reachable set of another convex shape, the same argument selects that set's tangency point. Its proof needs convexity, so it covers the one-query problem, not the hierarchical one. For one query, the proposition reads a training stall directly: a local minimum of training restricted to $\{\ell \leq R_i\}$ — network i, or a stage of training, with budget R_i — is the KKT point of Eq. (1) at R_i.

C The hierarchical model

Setting: § 3's tasks, prototypes, router and Eqs. (2)–(3).

For weights w_t summing to one with mixture $\bar{P} = \sum_t w_t P_t$,

$$\sum_t w_t \,\mathrm{KL}(P_t \| q) = \sum_t w_t \,\mathrm{KL}(P_t \| \bar{P}) + \mathrm{KL}(\bar{P} \| q) \tag{C.1}$$

(Cover & Thomas, 2006, Lemma 10.8.1). The first term does not involve q, so the mixture is the best single q.

The loss decomposes as

$$d(Q) = I(T;Y) - I(K;Y) + \sum_k m_k \,\mathrm{KL}(\bar{P}_k \| q_k), \tag{C.2}$$

where the information terms use the joint $\rho(t)\, p(k \mid t)\, P_t^\star(y)$. The first two terms are the outcome information the router discards (confusion). The last is each prototype's misfit to its pooled target (misfit).

Proof. Apply this identity (C.1) to each prototype's tasks with weights $p(t \mid k)$, using $\rho(t)\, p(k \mid t) = m_k\, p(t \mid k)$. The first terms sum to $H(Y \mid K) - H(Y \mid T) = I(T;Y) - I(K;Y)$, because $K - T - Y$ is a Markov chain. □

The information terms do not involve q_k. So prototype k enters the objective through $m_k \,\mathrm{KL}(\bar{P}_k \| q_k) + \lambda \,\mathrm{KL}(q_k \| \pi)$ alone, which gives Eq. (4)'s prototype update. Dividing by m_k: $q_k = Q^\star(\bar{P}_k; \lambda/m_k)$, the tilt (A.2) of π toward the pooled target at temperature λ/m_k.

Without the description term, the best prototypes are the pooled targets, $q_k = \bar{P}_k$. The problem then becomes the information bottleneck, with tasks as input and outcomes as the relevance variable (Tishby et al., 1999). Misfit vanishes, and (C.2) plus $\lambda I(T;K)$ is the bottleneck Lagrangian with $\beta = 1/\lambda$, up to scale and a constant.

The marginal minimizes $\sum_t \rho(t) \,\mathrm{KL}(p(\cdot \mid t) \| m)$ over m (Cover & Thomas, 2006) — (C.1) with routing distributions as the P_t. Free masses and masses tied to the router therefore share optima.

The router minimizes $\sum_k p(k \mid t) \,\mathrm{KL}(P_t^\star \| q_k) + \lambda \,\mathrm{KL}(p(\cdot \mid t) \| m)$, which is (A.1) with default m and value $-\mathrm{KL}(P_t^\star \| q_k)$; this gives Eq. (4)'s router. Routing costs at most $H(T)$ nats, which a Zipfian prior keeps finite. It also costs at most $H(K) \le \log K$ nats with K prototypes, the bound § 4 uses.

The alternation of Eq. (4) descends monotonically, and its limit points are stationary (Bertsekas, 1999, Prop. 2.7.1). Because the problem is not convex, a stationary point can be a local, not a global, minimum.

An analogue of Eq. (A.3) pins each prototype once the cells are fixed, since each prototype solves Eq. (1) for its cell at temperature λ/m_k. It does not pin the cells: at a jump, two optima with different cells tie.

D Formal treatment of Prediction 1

Setting: App. C's model, through Eq. (3) alone.

Power law, precise statement. Let $\mathcal{D}_n = \min_{q_1,\dots,q_n} \sum_t \rho(t) \min_{k \le n} \mathrm{KL}(P_t^\star \| q_k)$, the least loss n prototypes allow, cost aside. Assume every prototype other than π costs between c_- and c_+ nats of description. Then

$$\mathcal{D}_{1+R/c_-} \;\le\; \varepsilon(R) \;\le\; \mathcal{D}_{n(R)}, \qquad n(R) = \max\{n : \, nc_+ + \log(n+1) \le R\} \tag{D.1}$$

rounding R/c_- down. Whenever $\mathcal{D}_n = n^{-\gamma} L(n)$ with L slowly varying (Bingham et al., 1987), $\varepsilon(R)$ therefore decays as $R^{-\gamma}$ up to a slowly varying factor. With a common cost c, $\varepsilon(R) = \mathcal{D}_{R/c}\,(1 + o(1))$.

Lower bound. Routing costs are nonnegative, so a model with $\ell \le R$ has at most R/c_- prototypes besides π. Hence $d(Q) \ge \sum_t \rho(t) \min_k \mathrm{KL}(P_t^\star \| q_k) \ge \mathcal{D}_{1+R/c_-}$.

Upper bound. Route each task deterministically to its nearest prototype in an optimal n-codebook, with π as a free extra prototype. Routing then costs $H(K) \le \log(n+1)$ nats and description at most nc_+, and the loss is at most $\mathcal{D}_n$.

Exponent. $n(R) = R/c_+ - O(\log R)$, and regular variation gives $\mathcal{D}_{n+o(n)} = \mathcal{D}_n(1+o(1))$. Both bounds therefore share the exponent γ, and they coincide when $c_- = c_+$. □

Atomic tasks. The optimal n-codebook acquires the n most frequent atomic tasks (§ 3.1), so $\mathcal{D}_n = \Delta \sum_{t>n} \rho(t)$. With $\rho(t) = t^{-z}/\zeta(z)$, $\sum_{t>n} t^{-z} = n^{1-z}/(z-1) + O(n^{-z})$ gives $\varepsilon(R) = \frac{\Delta}{(z-1)\,\zeta(z)} \left(\frac{R}{c}\right)^{-(z-1)} (1+o(1))$.

Metric tasks. Suppose the targets $P_t^\star$ have a density on a d_T-dimensional smooth family, on which KL is locally half the Fisher quadratic form (Amari & Nagaoka, 2000). Then $\mathcal{D}_n \sim Cn^{-2/d_T}$ (Zador, 1982; Graf & Luschgy, 2000; Linder et al., 1999), and $\varepsilon \propto (R/c)^{-2/d_T}$. Prototypes act as piecewise-constant predictors. Piecewise-linear predictors on a data manifold give $4/d_T$ instead (Sharma & Kaplan, 2022), so the prototype family also sets the exponent.

E Formal treatment of Prediction 2

Setting: App. C, with (A.2), (C.1) and (C.2).

Plateau, precise statement. Let $s^\star = \sup_K I(K;Y)/I(T;K)$ over routers, the hypercontractivity coefficient of tasks and outcomes: the most outcome information any grouping of tasks buys per nat of routing. It lies between η^2 and 1 (Anantharam et al., 2013). Here $\eta^2 = \sup_\varphi \operatorname{Var} \mathbb{E}[\varphi(Y) \mid T] / \operatorname{Var} \varphi(Y)$ is the squared maximal correlation (Rényi, 1959): the largest share of an outcome feature's variance that the task explains. While $\lambda \ge s^\star$, one prototype serves every task: $Q^\star(\bar{P};\lambda)$. So one cell holds every task from $R = 0$, where every prototype sits at π, at least until λ falls to $s^\star$. Without the description term, the plateau is the information bottleneck's learnability threshold (Wu et al., 2019). The tasks of any one prototype likewise stay together while $\lambda \ge s_k^\star$, the same constant computed on those tasks; coarse distinctions come first when $s_k^\star < s^\star$ and the coarse split falls above $s_k^\star$, as it does without the description term.

Proof of the plateau. Take any model with routing information $r = I(T;K)$ and average prototype $\bar{q} = \sum_k m_k q_k$.

Loss bound. Eq. (C.2), $I(K;Y) \le s^\star r$, and joint convexity of KL (Cover & Thomas, 2006) give $d \ge I(T;Y) - s^\star r + \mathrm{KL}(\bar{P}\|\bar{q})$.

Resource bound. Since $m_k \le 1$, and by convexity, $\ell \ge r + \sum_k m_k \mathrm{KL}(q_k\|\pi) \ge r + \mathrm{KL}(\bar{q}\|\pi)$.

Combining. The two bounds give

$$d + \lambda\ell \ge I(T;Y) + (\lambda - s^\star)\, r + \mathrm{KL}(\bar{P}\|\bar{q}) + \lambda\, \mathrm{KL}(\bar{q}\|\pi). \tag{E.1}$$

For $\lambda \ge s^\star$, the right side is at least $I(T;Y) + \min_q[\mathrm{KL}(\bar{P}\|q) + \lambda\, \mathrm{KL}(q\|\pi)]$. That is the value one prototype at $Q^\star(\bar{P};\lambda)$ attains, since (C.1) gives one prototype the objective $I(T;Y) + \mathrm{KL}(\bar{P}\|q) + \lambda\, \mathrm{KL}(q\|\pi)$.

Uniqueness. For $\lambda > s^\star$, equality forces $r = 0$, so every task routes alike and each prototype's pooled target is $\bar{P}$. Let $V(\tau) = \min_q[\mathrm{KL}(\bar{P}\|q) + \tau\, \mathrm{KL}(q\|\pi)]$, which is strictly increasing when $\bar{P} \ne \pi$: its slope is $\mathrm{KL}(Q^\star(\bar{P};\tau)\|\pi) > 0$. Prototypes with masses m_k then cost at least $\sum_k m_k V(\lambda/m_k)$ beyond $I(T;Y)$, which exceeds $V(\lambda)$ whenever two prototypes carry mass. Every optimum therefore has one prototype.

Per prototype, and $s^\star \le 1$. On one prototype's tasks, the same argument charges routing at λ and description at λ/m_k, which gives $s_k^\star$. Data processing gives $s^\star \le 1$. □

Deterministic annealing starts at a temperature high enough that one prototype is optimal (App. E, plateau). It follows the stationary point as λ falls through the splits (Rose, 1998).

Separation. At every $\lambda > 0$, any two distinct prototypes of an optimum differ in some outcome's probability by at least $\lambda\, c\, q_{\min}$. Here $c > 0$ is the larger of their description costs $\mathrm{KL}(q\|\pi)$, and $q_{\min}$ is the smallest prototype probability.

Proof of separation. Let $q_j \neq q_k$ differ by at most δ in every outcome, with $c = \mathrm{KL}(q_k\|\pi)$ the larger description. Send k's tasks to q_j and drop q_k. This changes the objective in three ways:

Loss. Since $|\log a - \log b| \leq |a - b|/\min(a, b)$, the loss rises by at most $m_k\delta/q_{\min}$.

Routing. Routing information cannot rise, because the new route is a function of the old one (data processing; Cover & Thomas, 2006).

Description. Description falls by c.

An optimum admits no improvement, so $m_k\delta/q_{\min} \geq \lambda c$. Since $m_k \leq 1$, $\delta \geq \lambda c\, q_{\min}$. Prototypes solve Eq. (A.2) with a full-support default, so $q_{\min} > 0$. □

A split therefore cannot peel a prototype continuously away from a parent that costs $c > 0$: the two children differ by at least $\lambda c\, q_{\min}$ when they appear. A prototype at the default costs nothing, so its splits can be continuous. Two equally frequent tasks with $P^\star(\text{yes}) = 0.7$ and 0.3 pool to the uniform default, and their prototypes peel apart smoothly below $\lambda \approx 0.127$. The bottleneck alone would split them at $s^\star = (1 - 2 \cdot 0.3)^2 = 0.16$ (Anantharam et al., 2013).

Order (§ 3.1): an atomic task t with entry cost c_t enters once $\lambda < \lambda_t = \rho(t)\Delta_t/c_t$.

Proof of order. With each entry adding a fixed c_t, the tempered objective is $\sum_{t\notin A} \rho(t)\Delta_t + \lambda \sum_{t\in A} c_t$ plus a constant. This separates across tasks: t joins the acquired set A when $\rho(t)\Delta_t \geq \lambda c_t$. □

Order, details. The routing share of c_t is at most $\rho(t)\log(eM/\rho(t))$ nats, where M is the task mass not yet acquired. For rare tasks, description therefore dominates c_t. The routing bound follows from concavity of $x \mapsto -x\log x$: splitting mass ρ off a group of mass M raises $H(K)$ by at most $\rho\log(eM/\rho)$.

Frequency need not win. If costs rise with rarity, $c_t \propto t^b$, λ_t still falls with t, and acquiring the n most frequent tasks costs order n^{1+b} nats. The frontier then falls as $\varepsilon \propto R^{-(z-1)/(1+b)}$. A fact is a point-mass target that a model without it reads through its cell's prototype $\bar{P}_k$. If acquiring the fact means acquiring the point mass, App. A's point-mass cost with default $\bar{P}_k$ gives $c_t = \Delta_t = -\log \bar{P}_k(y_t)$, so $\lambda_t = \rho(t)$ whatever the fact's content: Eq. (6). A fact acquired as a partial tilt (A.2) would not cancel content.

F Formal treatment of Prediction 3

Setting: App. C's router, with $g_k = \mathrm{KL}(P^\star_{t^\circ}\|q_k) - \lambda\log m_k$ and $k^\circ = \arg\min_k g_k$ for a task t° without its own prototype (§ 3.1).

Proof of Prediction 3. At an optimum, each block of Eq. (4) minimizes the tempered objective given the others. The terms involving $p(\cdot \mid t^\circ)$ are

$$\rho(t^\circ)\Big[\sum_k p(k \mid t^\circ)\,\mathrm{KL}(P^\star_{t^\circ}\|q_k) + \lambda\,\mathrm{KL}(p(\cdot \mid t^\circ)\|m)\Big],$$

which the Gibbs tilt of m minimizes (A.1). Prototype k then gets $e^{-(g_k - g_{k^\circ})/\lambda}$ times the share of k°, so shares fall exponentially in their gap over λ. The prototypes are Eq. (4)'s third update. □

Size against fit. Suppose an unacquired task fits a small cell B ($m_B = 0.01$, KL 0.08) better than a large one A ($m_A = 0.30$, KL 0.10). At $\lambda = 0.01$, the prior's factor of 30 outweighs the evidence's $e^2 \approx 7.4$, and the answer is 80% A's. At $\lambda = 0.002$, the evidence e^{10} dominates, and the answer is 99.9% B's. The router weighs a cell's size against its fit, as Anderson's (1991) rational model of categorization does.

Four further predictions follow.

Content. A task outside the budget gets its cell's typical answer. When tasks share no structure, no acquired prototype fits the task, and it gets the generic prototype shared by the unacquired tasks. The answer is the winning cell's prototype $Q^\star(\bar{P}_k; \lambda/m_k)$ (App. C), which tilts π toward its members' pooled answers.

Frequency. A cell's prototype sits on Eq. (1)'s path at temperature λ/m_k, so rarer cells answer closer to the default. This matches the popularity effects in Kandpal et al. (2023) and Mallen et al. (2023), since a cell can hold a single fact.

Resolution. A cell stays together while $\lambda \geq s_k^\star$. Cells whose tasks reveal little about outcomes therefore keep a shared surrogate longest.

Rate. For atomic Zipfian tasks (Prediction 1), surrogates serve task mass $\varepsilon(R)/D \propto R^{-\alpha}$ with $D = \Delta$ and $\alpha = z - 1$ (App. D), so hallucination falls with the same exponent as loss. Kalai & Vempala (2024) derive a related floor from the mass of facts seen once.

G TRAINING DETAILS

The networks of Experiment 1 use rotary position embeddings (Su et al., 2024). Widths from 8 to 512 give $N = 1.86$k to 6.31M parameters. Each network trains in three seeds on $n = 2.1 \times 10^9$ tokens (262,144 steps of 128 documents). The optimizer is AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1; Loshchilov & Hutter, 2019), with a peak learning rate scaled with width, warmed up and then cosine-decayed.