Compression-Driven Abstraction Under Limited Inference: A Resource-Rational Account of Rules, Chunks, and Symmetries

Yuli Zhang (jp2022213694@qmul.ac.uk)[1]
[1]Queen Mary University of London, London, United Kingdom

Abstract

Humans can show abrupt gains in learning and problem solving, e.g., discovering a rule or forming a chunk. We propose a process-level account in which an agent selects among representation families by minimizing a two-part minimum description length (MDL) objective augmented with an explicit inference-cost penalty. Bounded inference is modeled as budgeted search with sparse top-k routing (limited consideration), yielding a retained-mass diagnostic α_k that predicts when truncation changes choices. The framework yields simple threshold conditions for when rules, macros, and symmetry codes become worthwhile, predicting step-like changes as experience accumulates under fixed resources. Minimal simulations reproduce key signatures, including compute–coverage tradeoffs and symmetry-threshold scaling with group size. Large language model (LLM) experiments further show a semantic–exact dissociation in symmetry learning and candidate-set bottlenecks consistent with limited consideration.

Keywords: minimum description length; resource rationality; abstraction; symmetry; chunking; in-context learning

Introduction

Humans often shift from surface-level, instance-based behavior to abstract and reusable representations. Classic examples include sudden *rule discovery* in concept learning and induction (Feldman, 2000; Goodman, Tenenbaum, et al., 2008; Shepard et al., 1961), *chunking* in skilled action and working memory (Chase & Simon, 1973; Cowan, 2001; Gobet et al., 2001; Mathy & Feldman, 2012; Miller, 1956), and exploiting *symmetries/invariances* to support efficient generalization (Feldman, 2016; Kemp & Tenenbaum, 2008; Wagemans, 1995). Such transitions are frequently experienced as abrupt improvements or "aha" moments, consistent with accounts that emphasize representational change (Kounios & Beeman, 2009; Ohlsson, 1992).

These phenomena suggest a shared computational tension: structured abstractions can yield large long-run gains in compression, planning, and generalization, but they require an up-front investment to learn and maintain. We present a unified account of these transitions as *resource-rational representation learning* (Anderson, 1990; Gershman et al., 2015; Lieder & Griffiths, 2020). Within our framework, an agent (i) performs inference under limited computation and (ii) chooses between competing hypothesis families by trading off representational complexity against data-fit and computation. When experience accumulates, the objective can exhibit an MDL "crossover" (Chater & Vitányi, 2003; Grünwald, 2007; Rissanen, 1978), producing threshold-like switches from memorization to abstractions.

To connect this abstract objective to recognizable cognitive signatures, we focus on three patterns that recur across classic literatures: (i) difficulty gradients and sudden *rule discovery* as structured hypotheses become worthwhile (Feldman, 2000; Shepard et al., 1961); (ii) *chunking* as compression that increases effective capacity under fixed budgets (Chase & Simon, 1973; Miller, 1956); and (iii) *symmetry/invariance* as orbit-level compression that can support recognition without enabling exact transform recovery (Kemp & Tenenbaum, 2008; Wagemans, 1995). Our simulations and LLM tasks instantiate these signatures in minimal, fully specified settings.

Contributions. (1) We formalize memorization, rules, macros/chunks, and symmetries under a single two-part MDL objective with an explicit compute cost. (2) We model bounded inference as sparse top-k routing (limited consideration) and provide a retained-mass diagnostic for when truncation is reliable. (3) We derive simple threshold conditions for the emergence of rules, macros/chunks, and symmetry-based compression, predicting step-like changes in behavior under fixed resources. (4) Minimal simulations validate compute–coverage tradeoffs, MDL-driven generalization jumps, macro-induced coverage jumps, and symmetry-threshold scaling. (5) LLM experiments validate symmetry scaling and candidate-set bottlenecks in a controlled computational substrate. Beyond standard MDL model selection, the key added commitments are the explicit coupling to process costs (λC) and a diagnostic (α_k) that links truncation to predictable failure modes and to measurable signatures under resource manipulations.

A resource-rational compression model

Tasks and hypothesis families

We consider supervised or goal-directed tasks defined by observations $D = \{(x_n, y_n)\}_{n=1}^{N}$. A hypothesis ("program") π maps inputs x to outputs y. Rather than committing to a fixed hypothesis space, we let Z denote a representation *family or coding scheme*, not a single stored table. Each choice of Z induces a hypothesis family:

- **Memorization** (instance table): store each (x_n, y_n), akin to exemplar-like storage (Love et al., 2004; Nosofsky, 1986).
- **Rules**: reusable symbolic constraints or compact programs (Feldman, 2000; Goodman, Tenenbaum, et al., 2008).

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

- **Macros/chunks**: compiled sub-derivations reusable across problems (chunking/temporal abstraction) (Gobet et al., 2001; Miller, 1956; Solway et al., 2014; Sutton et al., 1999).
- **Symmetries**: group actions identifying equivalent states/inputs and enabling orbit-level compression (Kemp & Tenenbaum, 2008; Tenenbaum et al., 2011; Wagemans, 1995).

The key question is *when* an agent should switch from a simpler but less general representation (memorization) to a more complex but more compressive one (rules/macros/symmetries).

Two-part MDL objective with compute

We model representation choice as minimizing a description length (Grünwald, 2007; Rissanen, 1978):

$$J(Z; D) = \underbrace{\ell(Z)}_{\text{representation cost}} + \underbrace{L(D \mid Z)}_{\text{data given } Z} + \lambda \underbrace{C(Z, D)}_{\text{inference cost}}, \tag{1}$$

where $\ell(Z)$ is the one-time cost of acquiring/maintaining Z (e.g., learning a rule library or macro), $L(D \mid Z)$ is the code length for data when using Z (e.g., residual exceptions or uncompressed items), and $C(Z, D)$ captures computation (e.g., search expansions), scaled by λ. MDL-style objectives are widely used for model selection and have been advocated as a criterion for cognitive model comparison (Lee, 2004; Myung & Pitt, 1997) and as a formalization of simplicity biases in cognition (Chater & Vitányi, 2003; Feldman, 2016).

Eq. (1) is intentionally abstract: it can be instantiated with standard two-part codes. In our simulations, each task fixes a concrete coding scheme and compute measure so that $\ell(Z)$ and $L(D \mid Z)$ share a common unit within that task (bits or "toy units"), and $C(Z, D)$ is implemented as a directly countable process cost (e.g., expansions or steps). Constants omitted from threshold derivations are held fixed across competing representations within each task and thus do not affect which Z is preferred. The central theoretical leverage is that switching between hypothesis families is governed by *MDL crossings* as experience (N, reuse M, or orbit count) grows. After a switch, the old representation need not be physically erased; the key prediction is that future storage or search cost grows more slowly, and in lossy settings a conversion cost can be added to $\ell(Z)$ or $C(Z, D)$.

Budgeted inference as limited consideration

Given a representation Z, inference selects hypotheses π to explain or solve tasks. We model a soft-MAP distribution

$$P(\pi \mid D, Z) \propto \exp\{-\beta\, E(\pi; D, Z)\}, \tag{2}$$

where E is an energy (negative log posterior plus optional search/effort penalties) and β controls decision noise. Limited consideration and truncated search have long been studied as models of bounded rationality (Gigerenzer & Goldstein, 1996; S. Russell & Wefald, 1991; Simon, 1955) and as rational approximations to Bayesian inference (Sanborn et al., 2010).

To model limited computation, we approximate inference by considering only the k best candidates at each routing step (top-k routing / sparse expansion). A useful runtime diagnostic is the *retained Gibbs mass*

$$\alpha_k = \frac{\sum_{i \in \text{top-}k} \exp(-\beta Q_i)}{\sum_j \exp(-\beta Q_j)}, \tag{3}$$

where Q_i are candidate costs. When $\alpha_k \approx 1$, truncation has negligible effect on soft decisions; when α_k is small, limited consideration can change outcomes. Although α_k is defined in terms of a candidate distribution, it yields behavioral predictions: conditions that lower α_k should increase sensitivity to additional deliberation (more compute) and increase instability under expanded candidate sets. For example, a practical proxy is *choice stability*: when α_k is low, allowing extra consideration (more time or a larger candidate window) should produce more revisions or slower, less confident decisions.

Online representation switching under limited inference

1. Observe new datum(s) and update per-Z sufficient statistics.
2. For each candidate representation Z: update $\widehat{L}_t(D \mid Z)$ and $\widehat{C}_t(Z, D)$; compute $\widehat{J}_t(Z)$.
3. Select Z_t (hard switch or soft mixture) and run top-k inference under Z_t; record α_k.
4. (Optional control) adapt compute: increase k or budget if α_k is low.

Algorithm 1: **A minimal process-level controller.** A lightweight online scheme implements Eq. (1) and couples representation choice to truncated inference via α_k.

A minimal online representation-switching algorithm

The objective in Eq. (1) can be implemented as a simple online controller rather than an offline comparison. At each update step t, maintain lightweight sufficient statistics for each candidate representation family Z (e.g., counts of unique facts, exceptions, macro uses, or orbit structure) and update plug-in estimates $\widehat{L}_t(D \mid Z)$ and $\widehat{C}_t(Z, D)$. Compute $\widehat{J}_t(Z) = \ell(Z) + \widehat{L}_t(D \mid Z) + \lambda\, \widehat{C}_t(Z, D)$ and select the active representation either by a hard switch $Z_t = \arg\min_Z \widehat{J}_t(Z)$ or by a soft mixture $P(Z_t \mid D) \propto \exp\{-\widehat{J}_t(Z)\}$. Inference within the chosen Z_t is then carried out by the top-k procedure; the diagnostic α_k can be used online to adapt compute (e.g., increase k when α_k is small). This minimal mechanism makes concrete the process-level claim: abrupt changes arise when estimated description-length curves cross, while resource limits (smaller k, smaller budgets) shift or smooth the crossing. *Hard vs soft switching.* Hard switching yields the sharpest threshold effects and serves as a clean limiting case; soft mixtures smooth transitions but preserve the qualitative prediction that representational change accelerates near MDL crossovers.

When do abstractions emerge? Threshold predictions

The MDL formulation yields simple, testable threshold conditions. We summarize them as qualitative predictions; the simulations in the next section instantiate the code lengths

explicitly. In all cases, tightening resource limits (smaller k, smaller budget B, or larger λ) shifts thresholds upward and can smooth otherwise sharp transitions (Lieder & Griffiths, 2020; Sanborn et al., 2010).

Rule emergence

Consider switching between a memorization code Z_M and a rule-based code Z_R that pays an up-front cost $\ell(r)$ to encode a rule r. Let
$Delta$ be the expected per-example savings in $L(D$
$midZ)$ when using r (relative to memorization). A rule becomes preferred when

$$NDelta \gtrsim \ell(r). \tag{4}$$

Thus, increasing rule complexity (larger $\ell(r)$) delays rule use, while increasing regularity (larger
$Delta$) accelerates it. This prediction connects to classic findings that concept learning difficulty tracks representational complexity (Feldman, 2000; Shepard et al., 1961) and to rational analyses of rule selection (Goodman, Tenenbaum, et al., 2008). *Compute-sensitive threshold.* If adopting r also reduces expected inference cost by
$Delta_C$ per example (e.g., fewer search expansions), then the crossover depends on both savings: $N($
$Delta + \lambda$
$Delta_C) \gtrsim \ell(r)$.

Macro/chunk emergence

A macro m (chunk) shortens solutions by s steps each time it is used, but has a one-time cost $\ell(m)$. Let M be the reuse count. A macro is worthwhile when

$$M\,s \gtrsim \ell(m). \tag{5}$$

Crucially, if each problem has a fixed step or time budget B, a macro can trigger a *coverage jump*: tasks that were previously unsolved within B become solvable. This is a normative expression of chunking/temporal abstraction as compression that increases effective capacity under constraints (Gobet et al., 2001; Mathy & Feldman, 2012; Miller, 1956; Sutton et al., 1999). *Compute-sensitive threshold.* When computation is explicitly penalized, macros can be preferred even earlier if they reduce inference cost (e.g., steps or expansions) by s_C per reuse: $M(s + \lambda s_C) \gtrsim \ell(m)$.

Symmetry emergence

Let a finite group Γ act on items so that observations fall into M orbits, with typical orbit size $|\Gamma|$. A symmetry-aware code pays ℓ_Γ to encode the group action, then stores orbit representatives (Kemp & Tenenbaum, 2008). In a semantic-use regime (only orbit identity matters), symmetry is favored when

$$M\,c \gtrsim \ell_\Gamma, \tag{6}$$

where c is the cost per uncompressed item. Because $N \approx |\Gamma|M$, the item threshold $N^\star$ scales sublinearly with $|\Gamma|$. In an exact-reconstruction regime (the specific group element must be recovered), an additional per-item cost shifts the threshold upward. This distinction mirrors cases where invariances support categorization/recognition without requiring recovery of the generative transformation (Feldman, 2016; Wagemans, 1995).

Computational simulations

We report minimal simulations designed to validate the qualitative signatures of the theory. All experiments are synthetic and require no human participants.

Top-k routing yields compute–coverage tradeoffs

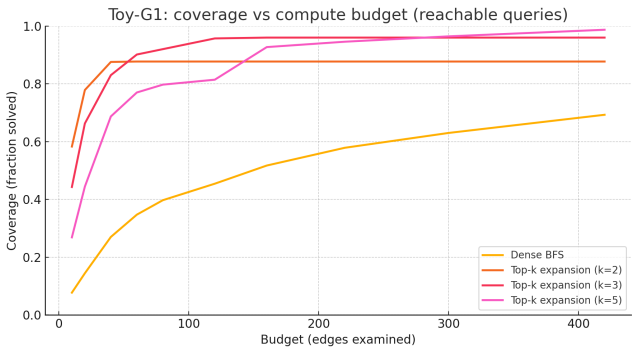
In a synthetic reachability task, inference corresponds to search over a latent directed acyclic graph under an edge-evaluation budget B (cf. budgeted heuristic search; Hart et al., 1968). We sample a random DAG (n=240 nodes; 900 source–target queries) and measure the fraction of queries solved within an edge-evaluation budget. Sparse top-k expansion achieves high solution coverage under limited B, approaching dense search as k increases; the diagnostic α_k tracks when truncation is reliable (Fig. 2). Supplementary experiment G5 couples this limited-consideration mechanism to an explicit compute penalty λ, producing threshold-like shifts in the preferred k and showing that α_k predicts reliability in the coupled setting.

MDL crossovers produce abrupt generalization

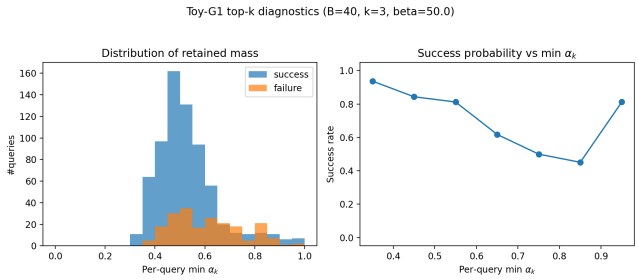
Model prediction experiment G3 instantiates an inductive setting with two competing explanation families: memorization versus a reusable rule. We use a commutativity rule in a small modular arithmetic toy domain (m=24; ℓ_{comm}=55; averaged over 120 random datasets). As the number of observations N increases, the two-part MDL costs cross (panel A), inducing a sharp shift in preferred explanation family. This shift yields a corresponding jump in held-out generalization/coverage (panel B), consistent with the model's prediction that representational switches can appear as step-like behavioral changes (Supplementary Fig. S1). Supplementary analyses replace hard switching with soft mixtures and show that the transition becomes smoother but preserves the qualitative crossover signature.

Macros induce coverage jumps under step budgets

Model prediction experiment G2 isolates the macro/chunk threshold in Eq. (5). Each problem requires executing a sequence of primitive steps; without a macro the shortest solution exceeds a step budget B, while introducing a reusable macro compresses the plan so it fits within B. We model macro adoption as MDL-driven: a one-time description cost is justified only after sufficient reuse. As reuse count M increases, macro use becomes preferred near the predicted threshold, producing an abrupt increase in success/coverage under the fixed budget (Supplementary Fig. S3).



(a) Coverage vs compute budget.



(b) α_k predicts truncation reliability.

Figure 2: **Model prediction experiment G1: top-k inference under budgets.** (A) Sparse top-k routing improves coverage under a fixed edge-evaluation budget B and approaches dense search as k increases. (B) The retained-mass diagnostic α_k tracks when truncation is reliable: higher per-query α_k corresponds to higher success probability.

Symmetry thresholds scale with group size

Model prediction experiment G4 constructs data organized into orbits under a finite group Γ. We compare a memory code to a symmetry-aware code in both semantic and exact-reconstruction regimes. In the plotted instance we set $(c, \ell_\Gamma) = (20, 2000)$ and vary $|\Gamma| \in \{2, 4, 8, 16, 32\}$. As predicted by Eq. (6), the orbit threshold $M^\star$ decreases quickly with $|\Gamma|$ in the semantic regime; requiring exact reconstruction shifts thresholds upward (Supplementary Fig. S4).

Worked examples: explicit code lengths for rules and symmetries

The threshold conditions in Eq. (4)–(6) are "ledger" inequalities, but the simulations above fully specify concrete codes. Here we write out two minimal worked examples (constants omitted).

Worked example (rules): G3 commutativity. Let items be ordered pairs (a, b) with $a, b \in \{2, \ldots, m-1\}$ and outputs $y = (a \times b) \bmod m$. A memorization code stores each observed oriented pair at cost c_{fact}, so $L_M(D) = N\, c_{\text{fact}}$. A commutativity representation pays $\ell(r) = \ell_{\text{comm}}$ once and stores each *unordered* observed pair $\{a, b\}$ at cost c_{fact}, giving $L_R(D \mid r) = |U|\, c_{\text{fact}}$ where $U = \{\{a, b\} : (a, b) \in D\}$. Because one stored unordered pair implies both orientations, the rule code yields higher expected test coverage at the same N (Supplementary Fig. S1B) and is preferred once $\ell_{\text{comm}} \lesssim (N - |U|)\, c_{\text{fact}}$.

Worked example (symmetries): G4 orbit code. Let a finite group Γ of size g act on items so that the dataset consists of M full orbits (so $N = gM$). A memory code stores each item at cost c, so $L_M(D) = Nc$. A symmetry-aware semantic code pays ℓ_Γ once and stores one representative per orbit, so $L_{\text{sym}}(D \mid \Gamma) = Mc$; preference requires $\ell_\Gamma \lesssim (g-1)Mc$, yielding $M^\star \approx \ell_\Gamma / ((g-1)c)$. For exact reconstruction, additionally encoding the group element index for each of the g items in an orbit costs $g \log_2 g$ bits per orbit, shifting the denominator to $(g-1)c - g \log_2 g$ and raising the threshold (Supplementary Fig. S4).

LLM experiments

We treat large language models (LLMs) as controllable computational subjects: each experiment provides only in-context evidence and requires a strict output format, with task-specific symbols randomized to reduce reliance on pretraining. All reported runs use greedy decoding (temperature = 0). To avoid confounds from changing chance levels across conditions (e.g., larger symmetry groups), we report *chance-corrected accuracy* $\tilde{a} = (a - a_0)/(1 - a_0)$ with binomial Wilson intervals. In this section, N refers to the in-context evidence size (number of labeled examples or labeled pairs) provided to the model. Full prompt templates, dataset generation details, and evaluation scripts are summarized in the supplementary materials. Unless otherwise noted, we report results for DeepSeek-V3.2 (32 trials per setting); a small sweep over other strong open-source models yielded the same qualitative signatures. A brief cross-model sweep (DeepSeek, Qwen, MiniMax) is reported in the supplement. Supplementary Fig. S5 summarizes the three experiments; Table 1 reports key numeric results.

Experiment A: Grid symmetries (semantic vs exact)

Design. We generate random 5×5 binary grids and apply a group action to form orbits under either rotations ($R4$) or rotations+reflections ($D8$). The *semantic* task asks the model to classify a grid by its orbit ID (equivalence class), while the *exact* task asks it to identify the specific group element mapping one grid to another. We vary the amount of training evidence (examples per orbit / transform-pairs) while holding the orbit structure fixed. *Notation.* In plots, N denotes the total number of labeled training items provided in-context (semantic) or the total number of labeled source–target pairs (exact). **Prediction.** Symmetry should help more in the semantic regime than in the exact-reconstruction regime, and

Experiment A: grids (R4 vs D8)		
Group	Semantic (N=18)	Exact (N=12)
R4	0.31 [0.23,0.40]	0.43 [0.37,0.50]
D8	0.73 [0.65,0.79]	0.20 [0.15,0.25]
Experiment B: sequences (C_g vs D_g)		
Group	Semantic	Exact (g=4/8)
C	1.00	0.86 [0.81,0.89] / 0.58 [0.53,0.64]
D	1.00	0.70 [0.65,0.75] / 0.31 [0.26,0.36]
Experiment C: candidate set bottleneck		
K	CorrAcc (N=1)	CorrAcc (N=6)
3	0.72 [0.47,0.87]	1.00 [0.84,1.00]
6	0.48 [0.27,0.66]	0.89 [0.71,0.96]
12	0.25 [0.10,0.44]	0.76 [0.58,0.88]
24	0.12 [0.03,0.29]	0.67 [0.49,0.81]

Table 1: **LLM summary statistics.** Cells report chance-corrected accuracy $\tilde{a} = (a - a_0)/(1 - a_0)$ with Wilson 95% intervals in brackets. Chance levels: semantic tasks $a_0 \approx 1/6$; ExpA exact: R4 = 1/4, D8 = 1/8; ExpB exact: $C_g = 1/g$, $D_g = 1/(2g)$; ExpC: $1/K$.

larger groups should make the semantic task easier relative to chance. **Result.** Both tasks are above chance, with a semantic–exact dissociation that grows with group size: for $D8$, semantic orbit classification is substantially higher than exact transform identification (Table 1; Supplementary Fig. S5A).

Experiment B: Sequence symmetries (semantic vs exact)

Design. We generate token sequences of length $g \in \{4, 6, 8\}$ with trial-unique symbols and consider cyclic (C_g) versus dihedral (D_g) symmetry actions. The semantic and exact tasks mirror Experiment A: orbit-ID classification versus identifying the transform label mapping source to target. **Prediction.** Semantic orbit classification should remain easy as g grows, while exact identification should degrade with g and be harder for D_g than C_g. **Result.** The model achieves near-ceiling semantic accuracy but shows a strong, graded drop in exact accuracy with increasing g, with $D_g < C_g$ at each g (Table 1; Supplementary Fig. S5B).

Experiment C: Candidate-set bottlenecks in rule identification

Design. Each trial samples a hidden substitution rule (a permutation over a small alphabet). The prompt provides N input–output examples and a list of K candidate rules that includes the true rule; the model must select the matching candidate index. **Prediction.** Increasing N should improve accuracy, while increasing K should decrease accuracy due to limited consideration. **Result.** Accuracy increases systematically with N and decreases with K, consistent with a candidate-set bottleneck (Table 1; Supplementary Fig. S5C). Interpreting $\tilde{a}$ under a simple limited-consideration model yields an *effective consideration size* $k_{\text{eff}} = 1 + \tilde{a}\,(K - 1)$, directly linking the curve to the theory's k parameter. Because increasing K also increases prompt length, a useful follow-up is a matched-length control (e.g., chunked candidate presentation) to further isolate limited consideration from context-length effects.

Related work

Our account sits at the intersection of (i) *resource-rational* theories that treat cognition as approximately optimal under computational costs, (ii) *compression/MDL* principles that formalize simplicity biases, and (iii) *structure learning* traditions explaining how people acquire rules, chunks, and invariances.

Resource rationality and approximate inference

A long-standing theme is that seemingly suboptimal behavior can be rational once one accounts for computational constraints (Gigerenzer & Goldstein, 1996; Payne et al., 1993; Simon, 1955, 1956). Modern *resource-rational* and *computational rationality* frameworks make these constraints explicit and connect them to process-level predictions (Gershman et al., 2015; Lieder & Griffiths, 2020). Related metareasoning perspectives formalize how an agent should allocate deliberation across alternatives when computation itself is costly (S. Russell & Wefald, 1991; S. J. Russell, 1997). A complementary line models probabilistic reasoning as *rational approximations* to ideal Bayesian inference (e.g., sampling-based or truncated-search procedures), where systematic deviations arise from cost-aware approximation rather than "irrationality" (Sanborn et al., 2010; Vul et al., 2014; Wilson et al., 2013). Our top-k consideration mechanism is a concrete instance of such truncation, and our retained-mass diagnostic provides a measurable link between resource limits and decision quality.

Compression, simplicity, and information constraints

Simplicity and compression have been proposed as unifying principles spanning perception, learning, and higher cognition (Chater & Vitányi, 2003; Feldman, 2016; M. Li & Vitányi, 2008; Solomonoff, 1964). The minimum description length (MDL) principle formalizes this idea as a tradeoff between model complexity and data fit (Grünwald, 2007; Rissanen, 1978) and has been used as a practical criterion for cognitive model evaluation (Lee, 2004; Myung & Pitt, 1997). From an information-theoretic angle, closely related objectives arise from information bottleneck and rate–distortion formulations of representation learning under constraints (Ortega & Braun, 2013; C. A. Sims, 2003; C. R. Sims, 2016; Tishby et al., 1999). Chunking can be interpreted as compression that increases effective capacity, linking descriptive complexity to behavior in working memory (Cowan, 2001; Mathy & Feldman, 2012; Miller, 1956). Our account extends these ideas by focusing on *discrete representation switching*: when a structured abstraction becomes worthwhile, the MDL objective can exhibit a crossing that yields threshold-like behavioral transitions.

Rules, chunks, and invariances as acquired structure

Rule-based concept learning has been studied across exemplar, prototype, and rule-based traditions (Feldman, 2000; Love et al., 2004; Nosofsky, 1986; Shepard et al., 1961) and can be modeled as rational hypothesis selection in structured rule spaces (Goodman, Tenenbaum, et al., 2008). Program induction approaches instantiate these ideas with compositional generative languages, enabling strong generalization from sparse data (Goodman, Mansinghka, et al., 2008; Griffiths et al., 2010; Lake et al., 2015). Related "library learning" systems (e.g., DreamCoder) learn reusable subroutines that amortize future inference, offering an algorithmic analogue of abstraction acquisition (Ellis et al., 2021). Chunking and macro-actions have also been formalized as temporal abstraction in planning and reinforcement learning (Barto & Mahadevan, 2003; Botvinick et al., 2009; Solway et al., 2014; Sutton et al., 1999). For invariances and symmetry, cognitive theories emphasize the discovery of structural form and the use of invariances to support generalization (Kemp & Tenenbaum, 2008; Tenenbaum et al., 2011; Wagemans, 1995). Our symmetry predictions make the additional distinction between semantic equivalence and exact reconstruction demands, which systematically alters when invariances pay off.

Phase-transition-like learning dynamics and insight

Threshold-like improvements are often associated with representational change (insight) in problem solving (Kounios & Beeman, 2014; Ohlsson, 1992, 2011). In machine learning, delayed-generalization phenomena ("grokking") show sharp shifts from memorization to generalization over training (Liu et al., 2022; Power et al., 2022). Our model provides a mechanism-level interpretation of such transitions as *MDL crossings* under limited consideration.

Large language models as computational subjects

Recent work argues that large language models (LLMs) can be used as controlled computational subjects to probe hypotheses about cognition and reasoning (Binz & Schulz, 2023; Niu et al., 2024; Qu et al., 2024). This perspective is especially natural for our framework because resource limits can be operationalized via token budgets, sampling budgets, tool-call budgets, or restricted candidate sets (De Sabbata et al., 2024; Y. Li et al., 2024; Wang et al., 2022; Wei et al., 2022). In-context learning analyses further suggest that LLM behavior can sometimes be interpreted as approximate Bayesian inference over tasks (Akyürek et al., 2023; von Oswald et al., 2023; Xie et al., 2021). We therefore view LLM experiments as complementary tests of our quantitative predictions about when abstractions become beneficial under explicit computational constraints.

Discussion

Scope and cognitive contact. CogSci routinely publishes computational and formal models, but reviewers expect a clear cognitive target and contact with established phenomena. A simulation-only submission is most compelling when it (i) proposes a process-level account of a recognizable signature (e.g., abrupt performance improvements interpreted as representational change (Kounios & Beeman, 2009; Ohlsson, 1992)), (ii) makes falsifiable predictions (e.g., threshold shifts under manipulations of complexity, regularity, or resource limits (Lieder & Griffiths, 2020)), and (iii) demonstrates robustness of the mechanism across task instantiations. Our goal here is the mechanism: MDL-driven representation choice under bounded inference yields unified thresholds for rules, chunks, and symmetries.

Contact with classic phenomena (no new data). The rule threshold in Eq. (4) offers a compact account of why concept learning difficulty tracks representational complexity (e.g., the Shepard–Hovland–Jenkins type structure and Boolean complexity gradients; Feldman, 2000; Shepard et al., 1961). The macro threshold in Eq. (5) provides a normative explanation of "sudden" gains from chunking when a compiled subroutine makes previously over-budget solutions feasible (Chase & Simon, 1973; Gobet et al., 2001; Miller, 1956). Finally, the semantic–exact symmetry distinction predicts when invariances should support recognition without enabling transform recovery, connecting to perceptual invariance phenomena (Wagemans, 1995).

LLMs as controllable "computational subjects." The LLM experiments in Section are not evidence of shared human mechanisms by themselves; they are controlled tests of the same resource manipulations in a tractable computational substrate (Binz & Schulz, 2023; Niu et al., 2024; Qu et al., 2024). Within our framework, limiting context or tool calls corresponds to restricting the effective k or budget B; increasing prompting complexity corresponds to increasing $\ell(Z)$; and enforced short deliberation corresponds to increasing the effective compute penalty λ (De Sabbata et al., 2024; Y. Li et al., 2024).

Limitations and next steps. The present simulations are intentionally minimal and do not fit any particular human dataset. A stronger follow-up would (i) scale to richer program spaces and learned libraries (Ellis et al., 2021), (ii) connect predictions to existing behavioral datasets in rule learning, chunking, and symmetry/invariance tasks (Chase & Simon, 1973; Shepard et al., 1961; Wagemans, 1995), and (iii) test resource-manipulated variants in controlled LLM agents as a complementary computational substrate (Wei et al., 2022; Xie et al., 2021).

References

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., & Zhou, D. (2023). What learning algorithm is in-context learning? investigations with linear models. *International Conference on Learning Representations*.

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates.

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, *13*(1–2), 41–77.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Botvinick, M. M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280. https://doi.org/10.1016/j.cognition.2008.08.011

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

De Sabbata, C. N., et al. (2024). Rational metareasoning for large language models. *arXiv preprint arXiv:2410.05563*.

Ellis, K., Wong, C., Nye, M., Sablon, A., Hewitt, L., Rasmussen, D., & Tenenbaum, J. B. (2021). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep program learning. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 3127–3138.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Feldman, J. (2016). The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(5), 330–340.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243.

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: A language for generative models. *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, 220–229.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring the laws of thought. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT Press.

Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, *4*(2), 100–107.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kounios, J., & Beeman, M. (2009). The AHA! moment: The cognitive neuroscience of insight. *Current Directions in Psychological Science*, *18*(4), 210–216.

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology*, *65*, 71–93.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lee, M. D. (2004). An efficient method for the minimum description length evaluation of cognitive models. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, 818–823.

Li, M., & Vitányi, P. M. B. (2008). *An introduction to kolmogorov complexity and its applications* (3rd ed.). Springer.

Li, Y., et al. (2024). Token-budget-aware reasoning for large language models. *arXiv preprint arXiv:2412.18547*.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

Liu, Z., et al. (2022). Towards understanding grokking: An empirical study of generalization in neural networks. *Advances in Neural Information Processing Systems*, *35*.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, *122*(3), 346–362.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

Niu, J., et al. (2024). Large language models and cognitive science: A comprehensive review. *arXiv preprint arXiv:2409.02387*.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. T. Keane & K. J. Gilhooly (Eds.), *Advances in the psychology of thinking* (pp. 1–44). Harvester Wheatsheaf.

Ohlsson, S. (2011). *Deep learning: How the mind overrides experience*. Cambridge University Press.

Ortega, P. A., & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A*, *469*(2153), 20120683. https://doi.org/10.1098/rspa.2012.0683

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets.

Qu, Y., et al. (2024). Promoting interactions between cognitive science and large language models. *Trends in Cognitive Sciences*.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Russell, S., & Wefald, E. (1991). *Do the right thing: Studies in limited rationality*. MIT Press.

Russell, S. J. (1997). Rationality and intelligence. *Artificial Intelligence*, *94*(1–2), 57–77.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational Bayesian inference. *Cognitive Science*, *34*(8), 128–142.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). *Learning and memorization of classifications*. Wiley.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, *50*(3), 665–690. https://doi.org/10.1016/S0304-3932(03)00029-1

Sims, C. R. (2016). Rate-distortion theory and human perception. *Cognition*, *152*, 181–198.

Solomonoff, R. J. (1964). A formal theory of inductive inference. part I. *Information and Control*, *7*(1), 1–22.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (2014). Optimal behavioral hierarchies in human planning. *PLoS Computational Biology*, *10*(8), e1003779.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1–2), 181–211.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., & Vladymyrov, M. (2023). Transformers learn in-context by gradient descent.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done: Optimal decisions from very few samples. *Cognitive Science*, 1–39. https://doi.org/10.1111/cogs.12101

Wagemans, J. (1995). Detection of visual symmetries. *Spatial Vision*, *9*(1), 9–32.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models.

Wilson, R. C., Nassar, M. R., & Gold, J. I. (2013). A mixture of delta-rules approximation to Bayesian inference in change-point problems. *PLoS Computational Biology*, *9*(7), e1003150.

Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2021). An explanation of in-context learning as implicit Bayesian inference.