Measuring the Unmeasurable: Unveiling Latent Cognitive Capabilities of LLM

Cui Danxin[2†], Sihang Jiang[1†], Keyi Wang[1], Zhiyi Duan[1], Yanghua Xiao[1*], Bi Yude[2*], Jiaqing Liang[1], Minggui He[3], Shimin Tao[3], Yilun Liu[3]

[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, China
[2] College of Foreign Languages and Literature, Fudan University, China
[3] Huawei, China
dxcui22@m.fudan.edu.cn, jiangsihang@fudan.edu.cn, shawyh@fudan.edu.cn, biyude@fudan.edu.cn

Abstract

As large language models (LLMs) are increasingly deployed in high-stakes domains such as education, healthcare, and law, accurately evaluating their nuanced reasoning process becomes essential to ensure their safety, reliability, and trustworthiness. However, most existing benchmarks evaluate LLMs at a coarse granularity. Current benchmarks lack a unified framework and rely on single-task datasets, overlooking the intermediate steps of complex reasoning. This results in redundant overlap across benchmarks, poor generalization to multifaceted real-world tasks, and underutilizes the rich reasoning traces generated by advanced LLMs.

Datasets — https://github.com/CogProbe/CogEval.git

Introduction

Large language models (LLMs) have made rapid advancements, achieving impressive performance across a wide range of standard evaluation tasks and demonstrating strong potential for real-world applications (Bubeck et al. 2023; Webb, Holyoak, and Lu 2023a). As their use expands into increasingly complex scenarios such as medical diagnosis and legal reasoning (Singhal et al. 2023; Griot et al. 2025), LLMs face a growing challenge: their performance in these high-stakes contexts is inconsistent, with models frequently making elementary logical errors despite high accuracy on existing benchmarks (Yang et al. 2023). This variation in performance highlights the need to develop a comprehensive model capability profile that identifies the strengths and weaknesses of different models, enabling targeted improvement and better application matching. Therefore, it is essential to design evaluation methods that not only measure overall performance but also focus on the reasoning steps involved in decision-making.

In recent years, several prominent benchmarks for evaluating large language models (LLMs) have emerged, including MMLU (Hendrycks et al. 2021), SciEval (Sun et al. 2024), and GPQA/SuperGPQA (Rein et al. 2023), which emphasize cross-disciplinary knowledge and advanced reasoning; HELM (Liang et al. 2022), BIG-bench (Srivastava et al. 2023), and PandaLM (Wang et al. 2024), which focus on comprehensive evaluations across multiple tasks and

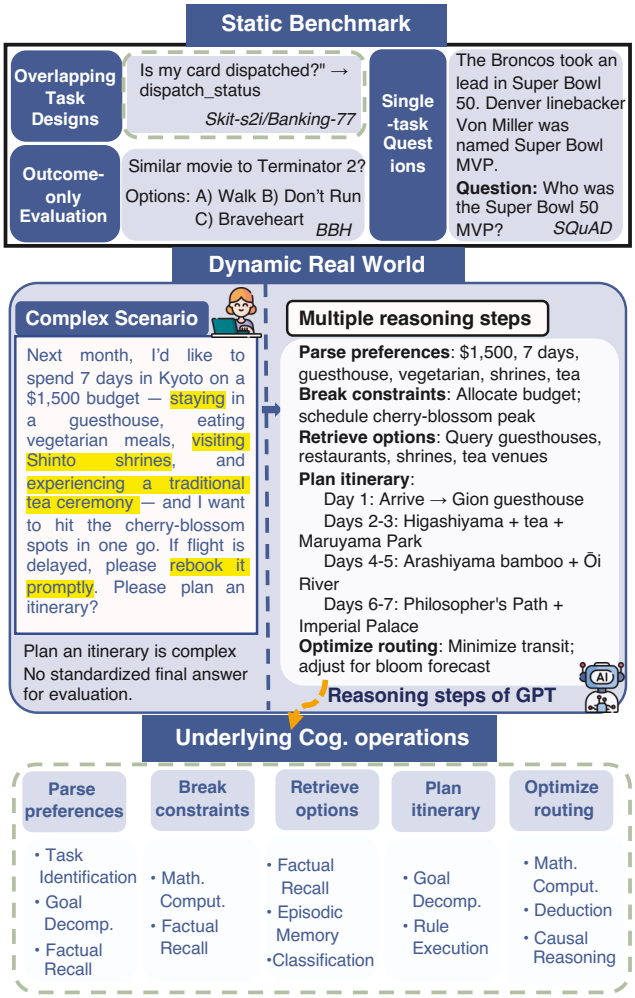

Figure 1: Existing benchmarks typically consist of overlapping, static single-task evaluations that ignore task interactions and intermediate reasoning process, whereas real-world scenarios involve interwoven tasks requiring complex, multi-step inference that can be decomposed into orthogonal, measurable components via a cognitive-theoretic framework.

[†]These authors contributed equally.
[*]Corresponding Authors

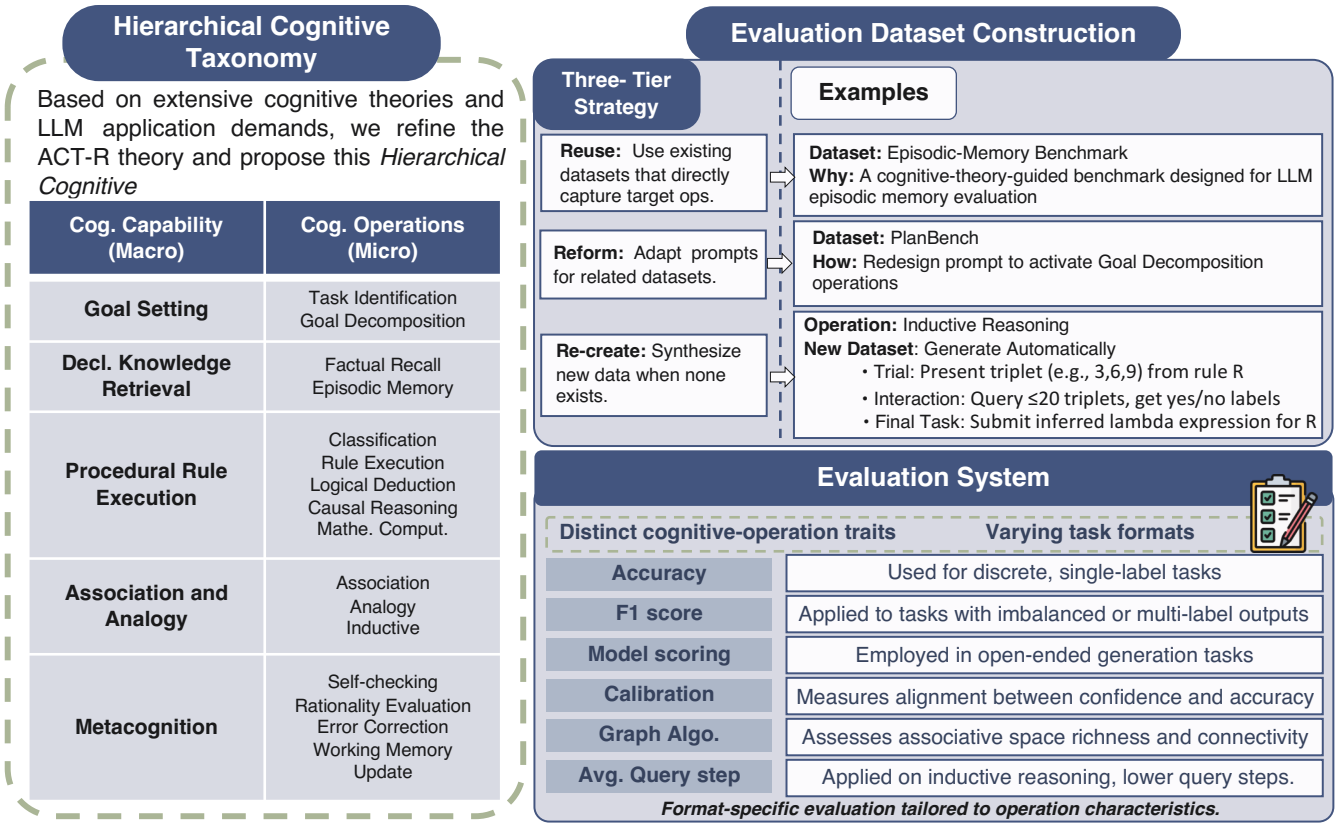

Figure 2: The framework of our benchmark design. We first establish a taxonomy containing five cognitive capabilities and 16 cognitive operations for complex task and reasoning process decomposition, then construct an evaluation dataset covering 16 operations, and finally propose a composite evaluation system tailored to operation characteristics.

metrics; and MM-Eval (Son et al. 2024), which targets multilingual fairness and impartiality. These benchmarks have substantially advanced the diversity and granularity of model assessment, providing rich insights into the capabilities of LLMs.

Despite these contributions, current benchmarks exhibit critical limitations, which constrain the deployment of LLMs in real-world applications. First, the single-task-centric design suffers from poor generalizability. With the rapid advancement of reasoning-intensive LLMs, models have become capable of supporting increasingly complex real-world applications, which typically require performing multiple interconnected tasks simultaneously. However, existing evaluation frameworks that focus solely on single-task assessments fail to effectively capture the interactions and synergies among these interconnected tasks. (Cao et al. 2025; Mondorf and Plank 2024b). Second, the absence of a unified theoretical framework leads to empirically crafted assessment tasks that often overlap, thereby impeding systematic and comprehensive capability assessment (Chang et al. 2024; Liang et al. 2022). Third, evaluating models solely based on final outcomes without examining intermediate reasoning steps neglects valuable reasoning information, which is essential for practitioners to gain deeper insights into a model's cognitive processes and identify potential reasoning flaws (Saparov and He 2023; Nguyen et al. 2024). Hence, severely constraining researchers' and practitioners' understanding and assessment of models' actual strengths and weaknesses, and preventing the creation of detailed capability profiles necessary for matching appropriate models to specific application scenarios.

Luckily, we can conceptualize the complex reasoning process as an information-processing system, allowing us to leverage *information processing theory* from cognitive psychology to break down complex tasks into discrete cognitive capabilities. Specifically, draw inspiration from the adaptive control of thought-rational cognitive architecture (ACT-R) (Anderson 1976) and complementary cognitive psychology theories, which offers a unified theoretical basis for modular and interactive cognitive processes, we propose a hierarchical cognitive evaluation taxonomy. This taxonomy comprises 5 macro cognitive-capability dimensions and 16 micro cognitive-operation dimensions. It disassembles the multi-step reasoning process required for complex task solving into components, each targeting distinct cognitive capabilities that are orthogonal and quantifiable.

To realize this taxonomy in practice, we develop the Cognitive Operations Probing Benchmark (*CogProbe*). And accordingly, we construct *CogEval* multilingual evaluation dataset and propose an evaluation system derived from cog-

nitive experimental paradigms that is suitable for measuring LLMs' cognitive capabilities.

Overall, our contributions are mainly fourfold:

- We present a systematic LLM evaluation framework grounded in information-processing theory from cognitive psychology, which decomposes problem-solving into constituent cognitive capabilities. Unlike prior work that focuses on single cognitive operations or proposes experience-driven taxonomies, our cognitively-grounded approach addresses key limitations of task-centric evaluation and provides a principled basis for assessing reasoning-intensive LLMs and informing optimal model selection for real-world applications.
- We propose a hierarchical cognitive evaluation taxonomy, which maps LLMs' latent cognitive capacities onto 16 cognitive operations, enabling fine-grained decoupling of complex reasoning processes;
- Based on our taxonomy, we introduce CogProbe, a multilingual diagnostic benchmark specifically designed to systematically evaluate the cognitive capabilities of LLMs. To operationalize CogProbe, we construct CogEval, a multilingual evaluation dataset carefully annotated with explicit cognitive operation demands. CogEval includes tasks sourced through a three-tiered approach: reusing existing LLM evaluation datasets, reforming tasks according to cognitive paradigms, and recreating novel tasks to specifically measure associative thinking, inductive reasoning and metacognitive operations.
- We conduct extensive experiments on state-of-the-art LLMs to evaluate their cognitive capabilities, locate cognitive strengths and weaknesses, and construct detailed capability profiles. We also compare cross-language performance. Our findings reveal substantial performance variations across cognitive dimensions, with notable deficiencies in metacognitive capabilities, and demonstrate that language also influences model performance.

Related Work

Benchmarking Mainstream evaluation benchmarks for large language models (LLMs), such as GLUE (Wang et al. 2019), SuperGLUE (Wang 2019), MMLU (Hendrycks et al. 2021), GSM8K (Cobbe et al. 2021), and recent comprehensive suites like BIG-Bench (Srivastava et al. 2023), HELM (Liang 2023), and FLASK (Ye et al. 2024), primarily adopt static, task-oriented evaluation methodologies. They predominantly rely on aggregate scoring of final outcomes, often overlooking the intermediate reasoning steps critical to problem-solving (Lanham 2023; Lyu 2023). As a result, such evaluations may reward pattern matching or memorization rather than genuine reasoning (Li 2023; Mondorf and Plank 2024a).

Process-oriented Evaluation Attempts Recent studies have begun to address this issue by explicitly evaluating intermediate reasoning processes. Notably, the emergence of chain-of-thought prompting (Wei 2022; Kojima 2022) inspired benchmarks such as Multi-LogiEval (Patel et al. 2024), and related verification methods (Lightman et al. 2023; Ling et al. 2023). Although these approaches have advanced the field by highlighting reasoning pathways, they remain largely empirical and lack systematic theoretical frameworks, resulting in fragmented and inconsistent evaluation criteria (Lanham 2023; Lyu 2023).

Cog. Capability (Macro)	**Cog. Operations (Micro)**
Goal Setting	Task Identification
	Goal Decomposition
Declarative Knowledge Retrieval	Factual Recall
	Episodic Memory
Procedural Rule Execution	Classification
	Rule Execution
	Logical Deduction
	Causal Reasoning
	Mathematical Computation
Association and Analogy	Association
	Analogy
	Inductive Reasoning
Metacognitive Monitoring	Self-checking
	Rationality Evaluation
	Error Correction
	Working Memory Update

Table 1: Hierarchical cognitive taxonomy for decomposing complex reasoning into observable, measurable components.

Cognitive Psychology in LLM Evaluation An emerging line of work applies cognitive psychology methods to analyze LLM behaviors (Binz and Schulz 2023; Webb, Holyoak, and Lu 2023b; Lampinen et al. 2024). Findings reveal human-like but fragile traits: GPT models show bounded working memory (Gong, Wan, and Wang 2024), limited theory-of-mind generalization (Ullman 2023), and heuristic biases in decision-making (Coda-Forno et al. 2024). CogBench integrates multiple such tasks into a broad benchmark (Coda-Forno et al. 2024), but its empirically-driven taxonomy lacks grounding in formal cognitive theory, relying on heuristics rather than a unified cognitive architecture. In contrast, CogProbe formalizes these insights through an ACT-R–inspired taxonomy, enabling systematic and interpretable cognitive evaluation.

Methodology

As shown in Fig. 2, we first decomposed the multi-step reasoning process required in solving complex task into discrete, observable, and measurable components, based on information processing theories, then construct an evaluation dataset, and finally propose an evaluation system tailored to cognitive operation characteristics.

Hierarchical Cognitive Taxonomy Design

In order to dissect complex tasks into orthogonal and measurable components, we resort to ACT-R architecture and its supplemental research by Salvucci and Cox (Anderson and Lebiere 1998; Salvucci and Anderson 2001; Cox, Oates,

Cog. Capability (Macro)	Cog. Operations (Micro)	Samples (En/Es/Cn)	Evaluation Metrics	Strategy
Goal Setting	Task Identification	300	Accuracy	Reform
	Goal Decomposition	285	Accuracy; F1	Reform
Decl. Knowledge Retrieval	Factual Recall	1000	Accuracy	Reuse
	Episodic Memory	200	Accuracy	Reuse
Procedural Rule Execution	Classification	300	Accuracy	Reform
	Rule Execution	180	Accuracy	Reform
	Logical Deduction	211	Accuracy	Reuse
	Causal Reasoning	300	Accuracy	Reuse
	Mathematical Computation	399	Model scoring	Reuse
Association and Analogy	Association	2000	Graph algorithm	Re-create
	Analogy	380	Accuracy; Model scoring	Reuse
	Inductive Reasoning	500	Accuracy; Avg. Query	Re-create
Metacognitive Monitoring	Self-checking	4055	Accuracy	Re-create
	Rationality Evaluation	4055	Calibration	Re-create
	Error Correction	4055	Accuracy	Re-create
	Working Memory Update	4055	Model scoring	Re-create

Table 2: The number of questions is based on the characteristics of cognitive operations. The data for the four metacognitive operations comes from the other 11 operations (except associative thinking). After answering, the model is prompted to reflect, provides a rationality evaluation (reasonable/irrational), a confidence score, and the reflection process. The proficiency of the four metacognitive operations is then calculated.

and Perlis 2011). ACT-R theory posits that task execution and problem-solving emerge from the coordinated activity of five independent cognitive modules; accordingly, it decomposes the process of solving complex tasks into modular information-processing components, which are orthogonal and operate in parallel (Fodor 1983; Anderson and Lebiere 1998; Taatgen and Anderson 2002). Therefore, assessing a model's performance across these five modules allows us to predict the types of task scenarios for which it is best suited. Next, because cognitive capabilities are latent and not directly measurable and each is instantiated via a set of corresponding micro-level cognitive operations (Anderson et al. 2004), we leverage additional cognitive theories and define specific operations for each of the five macro capabilities. Together, these two levels form our hierarchical macro–micro cognitive taxonomy, providing a principled, theory-driven basis for fine-grained evaluation (Anderson 2004; Salvucci and Anderson 2001; Cox, Oates, and Perlis 2011).

Evaluation Dataset Construction

We adopt a data construction strategy directly grounded in cognitive psychology theories and established experimental paradigms. Rather than simply aggregating existing datasets, we carefully select data and design prompt based on detailed cognitive-psychological definitions of each cognitive operation, as illustrated in Appendix. To achieve this, we propose a structured three-tier approach to dataset construction: **reuse**, **reform**, and **recreate**.

Reuse: Direct Mapping For micro-level operations explicitly captured by existing dataset, we directly integrate these tasks without alteration. Our selection process ensures that datasets precisely activate targeted cognitive operations through rigorous input-output analysis and core mechanism identification validated by cognitive psychology experts. Additionally, we verify that selected datasets exclude confounding factors from extraneous cognitive operations that could compromise measurement purity.

Reform: Data Adaption Certain datasets demand adaptation when their native prompts fail to elicit the intended cognitive operations or when the tasks, though conceptually related, are only indirectly aligned and therefore require transformation. Data adaptation is based on two complementary strategies: *prompt re-design*, which preserves the original task semantics while replacing native instructions with standardized, cognition-oriented prompts; *task transformation*, which reformulates only indirectly related tasks, for example, recasting PlanBench's path-planning problems as goal-decomposition exercises by automatically extracting their latent subgoals from the original questions.

Recreate: Data Creation For cognitive operations with scarce data, we synthesize new datasets by tailoring classical experimental paradigms from cognitive psychology for following dimensions (detailed description of these experimental paradigms shown in Appendix):

- **Inductive Reasoning** We adapt the Wason 2–4–6 rule-discovery paradigm (Wason 1960). Each trial presents three numbers from an unknown lambda function R. The model proposes up to 20 triplets, receiving "yes/no" feedback from a discriminator. Success rate and query count measure inductive accuracy and efficiency.
- **Associative Thinking** Drawing on semantic network methods (Kenett 2024), we present the model with 2,000 cue words and, at temperatures 0.3 and 0.8, elicit three rounds of ten lexical associates per cue. Temperatures of

0.3 and **0.8** were chosen to clearly distinguish deterministic from divergent associative behaviors, while avoiding overly restrictive (≤ 0.2) or excessively random (≥ 0.9) responses. We then compute pairwise symmetrical semantic similarities among all associates to construct an undirected graph for each temperature.

- **Metacognitive Operations:** Four metacognitive operations (Self-checking, Rationality Assessment, Error Correction, and Working Memory Update) share a common dataset drawn from our eleven micro-operation tasks (excluding associative thinking). After each initial response, the model receives a "confidence prompt" to rate its confidence and a "rationality prompt" to judge its reasoning. The model may then revise its answer. We compute calibration (confidence–accuracy correlation), reflection-driven accuracy change, and correction success rate to assess self-monitoring sensitivity and corrective effectiveness.

Multi-lingual Data Evolution Following the established framework, we first constructed English datasets and then generated corresponding Chinese and Spanish versions using Qwen-MT-Max model for translation.

Evaluation system

As shown in Tab. 2, our evaluation framework integrates accuracy, F1 score, model scoring, confidence scoring, and graph-theoretic metrics. Drawing from ACT-R theory, each cognitive operation engages distinct cognitive modules with unique underlying mechanisms. Thus, different metrics reflect fundamentally distinct cognitive constructs, consistent with measurement validity principles in psychometrics (Messick 1989). To capture operation-specific nuances, we introduce the following specialized metrics:

- **Graph Algorithm:** Associative thinking is difficult to capture using standardized metrics. Cognitive psychology research adopts following metrics for its measurement (Steyvers and Tenenbaum 2005; Kenett 2024) *Efficiency* (E) measures how rapidly information can traverse the semantic network, indicating the model's ability to jump between concepts. *Clustering coefficient* (C) quantifies the tendency of associates to form tightly-knit groups, reflecting local coherence. The *small-world index* (σ) compares E and C to a random network, revealing the balance between global reach and local density. *Community count* counts distinct semantic clusters, while *modularity* (Q) measures the strength of division into those clusters. *Semantic entropy* ($\bar{H}$) captures overall diversity of associations, with higher values denoting broader lexical exploration.
- **Average query:** Beyond accuracy, we compute the *mean query length*, i.e. the average number of queries required before correct hypothesis submission. A shorter mean indicates more efficient inductive inference.
- **Metacognitive operations profiling:** We adopt accuracy for Self-checking and Error Correction evaluation, calibration for Rationality Assessment and model scoring for Working Memory Update. The final performances of each of the four metacognitive operations are reported separately from the other twelve cognitive operations. Because metacognitive sensitivity depends on underlying task proficiency, observing metacognitive scores per operation also provides insight into the model's true competence in that operation.
- **Model scoring:** We use model scoring when correctness cannot be directly compared with gold answers. For **analogical reasoning**, the evaluator assesses logical coherence and similarity to gold-standard reasoning. For **mathematical reasoning**, it both the logical validity of intermediate steps and the correctness of the final numeric result relative to the gold answer.

Evaluation of the Benchmark

For cognitive-alignment assessment, each sample is annotated by four PhD-level cognitive-science experts independently to confirm that each prompt and item unambiguously invokes only the intended cognitive operation; pairwise Cohen's κ and Fleiss' κ reached **[0.76 ± 0.04]** and **[0.79]**, respectively ("substantial" agreement). Items with consensus from ≥ 3 annotators were retained. Non-English samples were translated using Qwen-MT-Max model and validated by human experts through structured sampling. We employed a *Minimum Quota + Proportional Allocation* strategy to ensure balanced representation across cognitive operations, yielding a representative 2 800-item sample. Bilingual experts evaluated adequacy and fluency on 1–5 Likert scales, achieving scores of **Adequacy = 4.83** and **Fluency = 4.79** with inter-rater reliability of $\alpha = 0.81$, indicating high agreement. This procedure ensured semantic fidelity and task consistency across languages.

Statistics of the Benchmark

Our final evaluation dataset comprises 30,330 instances of three languages (10,110 instances per language) covering 16 micro-level cognitive operations under five macro-capabilities. The number of samples is based on the characteristics of cognitive operations. For Factual Recall, we select a broad range of data due to the diversity of real-world knowledge. For Associative Thinking, we select 2000 cue word to ensure the coverage and representativeness of resulting semantic network. The four metacognitive operations share a single dataset (4055 samples per language), which is generated from the model's reflections on tasks involving the first 11 cognitive capabilities (excluding associative thinking).

Experiment

Models' performance across cognitive operations in the three languages is summarized in Tab. 3, *excluding* associative thinking and three metacognitive operations (self-checking, rationality evaluation and working memory update). Because metacognitive performance inherently depends on both the specific cognitive operation and the model's self-reflective capabilities, we report detailed metrics for these metacognitive dimensions separately in

Cog. Operations	Llama3-8B	Qwen3-32B (On)	Qwen3-32B (Off)	Qwen3-Plus (On)	Qwen3-Plus (Off)	GPT-4.1
Model Performance on English Data						
Task Identification	0.108 (↓0.027)	0.252 (↓0.015)	0.183 (↑0.020)	0.341 (↓0.012)	0.246 (↑0.007)	0.487 (↑0.027)
G. Decomp. (Acc/F1)	0 / 0	0 / 0.051 (↓0.002)	0 / 0.052 (↓0.001)	0 / 0.052 (↓0.002)	0 / 0.053 (↓0.002)	0 / 0.093 (↑0.005)
Factual Recall	0.830 (↓0.337)	0.775 (↓0.028)	0.410 (↑0.398)	0.834 (↓0.113)	0.570 (↑0.132)	0.881 (↓0.020)
Episodic Memory	0.830 (↓0.004)	0.952 (↓0.003)	0.918 (↓0.005)	0.978 (↑0.005)	0.961 (↓0.003)	0.970 (↑0.001)
Classification	0.097 (↓0.015)	0.863 (↓0.194)	0.817 (=)	0.881 (↓0.102)	0.839 (↑0.007)	0.247 (↑0.010)
Rule Execution	0.010 (↑0.023)	0.400 (↓0.270)	0.160 (↑0.013)	0.240 (=)	0.160 (=)	0.173 (↑0.073)
Logical Deduction	0.532 (↓0.201)	0.684 (↓0.050)	0.615 (↓0.007)	0.615 (=)	0.583 (↑0.007)	0.630 (↑0.036)
Causal Reasoning	0.746 (↓0.390)	0.840 (↓0.136)	0.810 (↓0.050)	0.863 (↓0.300)	0.816 (↓0.210)	0.810 (↑0.006)
Math. Comput.	0.541 (=)	0.954 (=)	0.954 (=)	0.965 (=)	0.972 (=)	0.964 (=)
Induct. (Acc/AvgQ)	0 / —	0.390 (↑0.024)/1.25	0.268 (=)/7.18	0.463 (↓0.169)/0.32	0.268 (↓0.065)/5.36	0.024 (=)/6
Analogy	0.300 (↑0.070)	0.676 (↓0.079)	0.620 (↑0.013)	0.676 (↓0.079)	0.625 (↑0.021)	0.520 (↑0.030)
Model Performance on Chinese Data						
Task Identification	0.037 (↓0.002)	0.256 (↓0.008)	0.220 (↑0.010)	0.248 (↓0.007)	0.239 (↑0.009)	0.354 (↑0.012)
G. Decomp. (Acc/F1)	0 / 0	0 / 0.052 (↓0.002)	0 / 0.052 (↓0.002)	0 / 0.050 (↓0.005)	0 / 0.052 (↓0.006)	0 / 0.106 (↓0.002)
Factual Recall	0.652 (↓0.193)	0.830 (↓0.143)	0.763 (↓0.043)	0.843 (↓0.173)	0.797 (↓0.120)	0.800 (↑0.067)
Episodic Memory	0.773 (↑0.045)	0.879 (↓0.003)	0.791 (↓0.008)	0.924 (↑0.007)	0.928 (↑0.008)	0.960 (↑0.015)
Classification	0.175 (↑0.026)	0.739 (↓0.102)	0.698 (↓0.006)	0.847 (↓0.007)	0.809 (↑0.024)	0.856 (↑0.021)
Rule Execution	0.037 (↓0.007)	0.267 (↓0.193)	0.267 (=)	0.267 (↓0.013)	0.193 (↓0.126)	0.160 (↑0.093)
Logical Deduction	0.439 (↑0.072)	0.640 (↓0.100)	0.540 (↓0.029)	0.568 (↓0.036)	0.518 (↑0.094)	0.606 (=)
Causal Reasoning	0.700 (↓0.037)	0.830 (↓0.143)	0.763 (↓0.043)	0.843 (↓0.067)	0.797 (↓0.12)	0.837 (↓0.127)
Math. Comput.	0.541 (=)	0.937 (=)	0.937 (=)	0.939 (=)	0.944 (=)	0.959 (=)
Induct. (Acc/AvgQ)	0 / —	0.222 (=)/1	0.073 (=)/15.67	0.342 (=)/1	0.171 (↑0.033)/3.714	0 / —
Analogy	0.459 (↑0.030)	0.739 (↓0.021)	0.654 (↑0.011)	0.746 (↓0.105)	0.664 (=)	0.551 (↑0.090)
Model Performance on Spanish Data						
Task Identification	0.142 (↓0.006)	0.220 (↓0.007)	0.174 (↓0.001)	0.246 (↓0.007)	0.203 (↑0.005)	0.339 (↓0.002)
G. Decomp. (Acc/F1)	0 / 0	0 / 0.052 (↓0.002)	0 / 0.051 (↓0.001)	0 / 0.052 (↓0.004)	0 / 0.055 (↓0.002)	0 / 0.083 (↑0.020)
Factual Recall	0.493 (↓0.384)	0.834 (↓0.113)	0.548 (↑0.172)	0.845 (↓0.176)	0.735 (↓0.225)	0.861 (↓0.092)
Episodic Memory	0.709 (↓0.058)	0.905 (↓0.008)	0.848 (=)	0.906 (↑0.002)	0.893 (↑0.095)	0.912 (↑0.051)
Classification	0.170 (↓0.026)	0.728 (↓0.085)	0.679 (↓0.010)	0.832 (↓0.076)	0.762 (↓0.048)	0.866 (↓0.031)
Rule Execution	0.003 (=)	0.004 (↓0.004)	0.0006 (=)	0.060 (↑0.002)	0.026 (↑0.020)	0.080 (↑0.040)
Logical Deduction	0.503 (↑0.010)	0.705 (↓0.079)	0.561 (↑0.079)	0.576 (↓0.014)	0.568 (=)	0.670 (↓0.043)
Causal Reasoning	0.743 (↓0.050)	0.837 (↓0.143)	0.813 (↓0.100)	0.840 (↓0.280)	0.827 (↓0.243)	0.810 (↓0.197)
Math. Comput.	0.607 (=)	0.942 (=)	0.937 (=)	0.949 (=)	0.959 (=)	0.946 (=)
Induct. (Acc/AvgQ)	0 / —	0.243 (↑0.048)/0.3	0.244 (=)/5.8	0.463 (=)/0.16	0.244 (↓0.041)/7.5	0 / —
Analogy	0.471 (↓0.003)	0.620 (↓0.002)	0.564 (↓0.002)	0.675 (↓0.106)	0.581 (↓0.005)	0.480 (↑0.080)

Table 3: Model performance across languages and cognitive operations. Each cell shows the baseline score followed by the change after reflection in parentheses; upward arrows (↑) indicate improvement, downward arrows (↓) indicate degradation, and (=) denotes unchanged. For **Induction**, cells report *accuracy / average queries*, and for **Goal Decomposition**, *accuracy / F1*. Full model details are provided in the Appendix.

the Appendix. The performance changes shown in Table 3 specifically represent models' metacognitive error-correction ability—i.e., their effectiveness in revising incorrect responses after explicit reflection. Associative thinking performance is quantified via graph-based metrics (efficiency, clustering, modularity) extracted from semantic networks generated by each model, and is likewise reported separately in the Appendix.

Evaluated Models We evaluate a total of 4 models ranging from their model size and supported context length. These models are categorized according to their model size. To assess the influence of deep reasoning capabilities on model performance, we conducted controlled comparative experiments using Qwen3-Plus and Qwen3-32B models, evaluating their performance with deep reasoning enabled and disabled.

Observation 1: Current LLMs fail to excel consistently across cognitive dimensions. Current large language models excel primarily in memory-driven tasks (e.g., *Factual Recall*, *Episodic Memory*), yet their performance significantly deteriorates on complex reasoning and induction tasks. Reflection-induced fluctuations further reveal inherent cognitive brittleness, highlighting substantial room for improvement in achieving balanced cognitive robustness.

Observation 2: Human cognitive difficulty hierarchies similarly apply to large language models. Bloom's Taxonomy posits six ascending cognitive levels—*Knowledge* (remembering facts), *Comprehension* (understanding meaning), *Application* (using information), *Analysis* (breaking down information), *Synthesis* (combining elements), and *Evaluation* (making judgments) (Anderson and Krathwohl 2001). LLM scores mirror this hierarchy: they excel at

lower-order levels (Knowledge/Comprehension), decline on higher-order reasoning, and perform worst on metacognition. All models score above 0.65 in *Factual Recall* and *Episodic Memory*, yet accuracy drops sharply for *Rule Execution*, *Logical Deduction*, and especially *Induction*. Reflection often decreases scores for smaller models (LLaMA3-8B-Instruct, Qwen3-32B), whereas GPT-4.1 alone maintains—or modestly improves—self-regulation. Hence, current LLMs have largely mastered basic recall but remain limited in higher-order reasoning and self-monitoring.

Observation 3: "Think-Mode On" Yields Large but Unstable Gains. For both QWEN3-32B and QWEN3-PLUS, enabling the internal *think mode* (explicit reasoning traces) significantly boosts performance across most cognitive operations, e.g., English *Rule Execution* improves notably from 0.16 to 0.40 and *Induction* from 0.27 to 0.46. However, these same models experience pronounced negative post-reflection deltas, such as a substantial -0.27 drop for Qwen3-32B (On) in *Rule Execution* and a -0.17 loss for Qwen3-Plus (On) in *Induction*.

We hypothesize that this instability arises because the explicit reasoning required by think mode exposes underlying inconsistencies in the models' representations. Instead of mastering symbolic reasoning, these models depend on brittle, surface-level statistical associations, which collapse during critical reflection, resulting in accuracy declines. For instance, in the *Inductive Reasoning* task, when "think mode" is off, both QWEN3-32B and QWEN3-PLUS systematically perform multiple queries to reach correct answers (average queries typically range between 3 and 8, with QWEN3-32B requiring as many as 15.67 queries in Chinese). However, with think mode activated, the models frequently guess immediately (average queries ≤ 1 across all languages), indicating a shift from inductive reasoning toward superficial pattern matching.

Observation 4: Metacognition is the key for diagnosing genuine capability. Metacognition is the capacity to monitor and critically evaluate one's own cognitive processes (Nelson and Narens 1990). It distinguishes genuine conceptual understanding from superficial pattern matching (Flavell 1979). Models relying on shallow statistical associations often show performance declines when explicitly prompted to reflect, misjudging their earlier correct answers as errors. Conversely, stable or improved accuracy after reflection signals robust internal representations and genuine mastery. Here is an example from inductive reasoning of Qwen3-plus:

Initially, given the numeric example $(x, y, z) = (47, 12, 59)$, the model incorrectly proposed a broad rule:

$$\text{lambda } x, y, z : (x + y == z) \text{ or } (x == y + z)$$

Upon explicit reflection, the model then correctly revised its answer:

$$\text{lambda } x, y, z : x + y == z$$

This correction clearly demonstrates the model's ability to critically reassess and rectify its initial erroneous response upon reflection.

Lang.	Temp.	E	C	σ_E	Q	Comm.	Avg. Ent.
EN	0.3	0.400	0.107	1.30	0.941	29	1.74
	0.8	0.426	0.189	1.42	0.941	29	1.96
ES	0.3	0.221	0.490	9.49	0.831	32	1.82
	0.8	0.209	0.430	12.48	0.838	25	2.06
CN	0.3	0.273	0.426	8.61	0.768	22	1.95
	0.8	0.387	0.516	1.72	0.335	19	2.29

Table 4: Semantic network metrics for associative thinking assessment using GPT-4.1.

Observation 5: Cross-linguistic differences primarily reflect training-data exposure rather than linguistic structure, except in associative thinking. Our results suggest that performance differences across languages primarily stem from disparities in training-data exposure, rather than intrinsic linguistic features. In most cognitive operations (e.g., *Factual Recall*, *Logical Deduction*, *Rule Execution*), performance differences between Chinese and Spanish remain minimal, both trailing English, indicating that data quantity and quality dominate model generalization capabilities rather than language structure per se. However, a clear exception arises in *Associative Thinking* (Table 4), where intrinsic linguistic structures strongly influence cognitive performance. Specifically, due to the semantic richness inherent in logographic characters, Chinese semantic network achieves the "small-world sweet spot"—balanced efficiency ($E \approx 0.39$) and clustering ($C \approx 0.52$) with moderate modularity ($Q = 0.34$)—closely matching human high-creativity patterns (Kenett 2024). English output is fast but shallow (high E, low C); Spanish is locally rich but globally fragmented (high C, low E), and both exhibit excessive modularity. These phenomena align well with cognitive linguistic theories, suggesting that language structures and cultural experiences inherently shape cognitive and associative processes differently across languages (Whorf 1956; Lakoff and Johnson 1980).

Conclusion

We propose a cognition-centric evaluation framework utilizing curated datasets and multidimensional metrics to profile large language models across tasks and languages. Experiments reveal that **(i)** models excel at memory-based tasks but struggle with higher-order reasoning, validating our cognitive hierarchy aligned with Bloom's taxonomy; **(ii)** enabling explicit reasoning ("thinking mode") improves accuracy but introduces volatility, particularly in smaller models, presenting a trade-off favorable in step-by-step reasoning tasks; **(iii)** metacognition critically differentiates genuine cognitive competence from superficial correctness; and **(iv)** cognitive performance across evaluated high-resource languages primarily reflects training data exposure, though associative thinking uniquely mirrors language-specific structures. These insights highlight the necessity of detailed cognitive profiling for effective model selection, deployment, and targeted improvements.

References

Anderson, J. R. 1976. *Language, Memory, and Thought*. Erlbaum.

Anderson, J. R. 2004. *Cognitive psychology and its implications*. New York, NY: Worth Publishers.

Anderson, J. R.; Bothell, D.; Byrne, M. D.; Douglass, S.; Lebiere, C.; and Qin, Y. 2004. An integrated theory of the mind. *Psychological Review*, 111(4): 1036–1060.

Anderson, J. R.; and Lebiere, C. 1998. *The Atomic Components of Thought*. Erlbaum.

Anderson, L. W.; and Krathwohl, D. R. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.

Binz, M.; and Schulz, E. 2023. Using Cognitive Psychology to Understand GPT-3. *Proceedings of the National Academy of Sciences (PNAS)*, 120(6): e2218523120.

Bubeck, S.; et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Cao, Y.; Hong, S.; Li, X.; Ying, J.; Ma, Y.; Liang, H.; Liu, Y.; Yao, Z.; Wang, X.; Huang, D.; Zhang, W.; Huang, L.; Chen, M.; Hou, L.; Sun, Q.; Ma, X.; Wu, Z.; Kan, M.-Y.; Lo, D.; Zhang, Q.; Ji, H.; Jiang, J.; Li, J.; Sun, A.; Huang, X.; Chua, T.-S.; and Jiang, Y.-G. 2025. Toward Generalizable Evaluation in the LLM Era: A Survey Beyond Benchmarks. arXiv:2504.18838.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3): 1–45.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, ; Schulman, J.; Hilton, J.; Nakano, R.; Hesse, C.; et al. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

Coda-Forno, J.; Binz, M.; Wang, J. X.; and Schulz, E. 2024. CogBench: a large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Cox, M. T.; Oates, T.; and Perlis, D. 2011. Toward an Integrated Metacognitive Architecture. *AAAI Technical Report*, FS-11-01: 74–81.

Flavell, J. H. 1979. Metacognition and Cognitive Monitoring: A New Area of Cognitive-Development Inquiry. *American Psychologist*, 34(10): 906–911.

Fodor, J. A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Gong, D.; Wan, X.; and Wang, D. 2024. Working memory capacity of ChatGPT: an empirical study. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press. ISBN 978-1-57735-887-9.

Griot, M.; Hemptinne, C.; Vanderdonckt, J.; Yuksel, D.; and et al. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16: 642.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Kenett, Y. N. 2024. The Role of Knowledge in Creative Thinking. *Creativity Research Journal*.

Kojima, T. e. a. 2022. Large Language Models are Zero-Shot Reasoners. *arXiv:2205.11916*.

Lakoff, G.; and Johnson, M. 1980. *Metaphors We Live By*. University of Chicago Press.

Lampinen, A. K.; Dasgupta, I.; Chan, S. C. Y.; Sheahan, H. R.; Creswell, A.; Kumaran, D.; McClelland, J. L.; and Hill, F. 2024. Language Models, like Humans, Show Content Effects on Reasoning Tasks. *PNAS Nexus*, 3(7).

Lanham, T. e. a. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv:2307.13702*.

Li, Y. e. a. 2023. Making Language Models Better Reasoners with Step-Aware Verifier. In *ACL*.

Liang, P.; et al. 2022. Holistic Evaluation of Language Models. *arXiv preprint arXiv:2211.09110*.

Liang, P. e. a. 2023. Holistic Evaluation of Language Models. *TMLR*.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's Verify Step by Step. arXiv:2305.20050.

Ling, Z.; Fang, Y.; Li, X.; Huang, Z.; Lee, M.; Memisevic, R.; and Su, H. 2023. Deductive verification of chain-of-thought reasoning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Lyu, Q. e. a. 2023. Faithful Chain-of-Thought Reasoning. In *IJCNLP*.

Messick, S. 1989. Validity. In Linn, R. L., ed., *Educational measurement*, 13–103. New York: Macmillan, 3 edition.

Mondorf, P.; and Plank, B. 2024a. Beyond Accuracy: Evaluating Reasoning Behavior of LLMs. *COLM*.

Mondorf, P.; and Plank, B. 2024b. Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models – A Survey. arXiv:2404.01869.

Nelson, T. O.; and Narens, L. 1990. Metamemory: A Theoretical Framework and New Findings. In Bower, G. H., ed., *The Psychology of Learning and Motivation*, volume 26, 125–173. New York: Academic Press.

Nguyen, M.-V.; Luo, L.; Shiri, F.; Phung, D.; Li, Y.-F.; Vu, T.-T.; and Haffari, G. 2024. Direct Evaluation of Chain-of-Thought in Multi-hop Reasoning with Knowledge Graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2862–2883. Bangkok, Thailand: Association for Computational Linguistics.

Patel, N.; Kulkarni, M.; Parmar, M.; Budhiraja, A.; Nakamura, M.; Varshney, N.; and Shah, C. B. 2024. MultiLogiEval: Evaluating Multi-Step Logical Reasoning in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 18889–18907. Association for Computational Linguistics.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *ArXiv*, abs/2311.12022.

Salvucci, D. D.; and Anderson, J. R. 2001. Integrating analogical mapping and general problem solving: The path-mapping theory. *Cognitive Science*, 25(1): 67–110.

Saparov, A.; and He, H. 2023. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. In *International Conference on Learning Representations (ICLR)*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S. R.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620: 172–180.

Son, G.; Yoon, D.; Suk, J.; and et al. 2024. MM-Eval: A Multilingual Meta-Evaluation Benchmark for LLM-as-a-Judge and Reward Models. *arXiv preprint arXiv:2410.17578*.

Srivastava, A.; et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Steyvers, M.; and Tenenbaum, J. B. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1): 41–78.

Sun, L.; Han, Y.; Zhao, Z.; Ma, D.; Shen, Z.; Chen, B.; Chen, L.; and Yu, K. 2024. SciEval: a multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press. ISBN 978-1-57735-887-9.

Taatgen, N. A.; and Anderson, J. R. 2002. How cognitive architectures have influenced cognitive models: A retrospective on EPIC, ACT-R, and SOAR. *Handbook of Cognitive Science: An Embodied Approach*, 103–143.

Ullman, T. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv:2302.08399.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*.

Wang, A. e. a. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *NeurIPS*.

Wang, Y.; Yu, Z.; Zeng, Z.; Yang, L.; Wang, C.; Chen, H.; Jiang, C.; Xie, R.; Wang, J.; Xie, X.; Ye, W.; Zhang, S.; and Zhang, Y. 2024. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. arXiv:2306.05087.

Wason, P. C. 1960. On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology*, 12(3): 129–140.

Webb, T.; Holyoak, K. J.; and Lu, H. 2023a. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(8): 1088–1101.

Webb, T.; Holyoak, K. J.; and Lu, H. 2023b. Emergent Analogical Reasoning in Large Language Models. *Nature Human Behaviour*, 7(9): 1526–1541.

Wei, J. e. a. 2022. Chain-of-Thought Prompting Elicits Reasoning in LLMs. *NeurIPS*.

Whorf, B. L. 1956. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press.

Yang, T.; Jin, Q.; Li, H.; Yu, Y.; Wu, H.; Wang, H.; Chen, K.; Yang, Y.; Ding, W.; Han, X.; et al. 2023. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *IEEE Transactions on Knowledge and Data Engineering*.

Ye, S.; Kim, D.; Kim, S.; Hwang, H.; Kim, S.; Jo, Y.; Thorne, J.; Kim, J.; and Seo, M. 2024. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. In *The Twelfth International Conference on Learning Representations*.