Detailed Definitions of Cognitive Operations for Prompt Design and Evaluation Metrics Selection

The 16 cognitive operations selected in our study are directly grounded in established cognitive psychology theories and classical experimental paradigms. Each cognitive operation corresponds explicitly to a specific cognitive theoretical framework or well-documented experimental paradigm from cognitive psychology literature. The detailed correspondence between each cognitive operation and its theoretical foundations is summarized in Tab. 8 and Tab. 9.

Based on these researches, we provide concise psychological definitions for each of the 16 cognitive operations used in our evaluation framework. We also explain the rationale behind selecting specific evaluation metrics tailored to each operation (e.g., graph-based algorithms for associative thinking and calibration metrics for metacognition), grounded in relevant cognitive psychology theories.

Concise Definitions of Cognitive Operations

0.1 **Task Identification:** Identifying the nature, scope, and constraints of a presented task based on initial instructions or situational cues.

0.2 **Goal Decomposition:** Breaking down complex objectives into manageable, actionable sub-goals, facilitating efficient task completion.

0.3 **Factual Recall:** Retrieving previously learned factual information from long-term memory without external cues.

0.4 **Episodic Memory:** Recalling context-specific past experiences or events, including associated temporal and spatial information.

0.5 **Classification:** Categorizing stimuli or concepts into predefined groups based on shared properties or rules.

0.6 **Rule Execution:** Applying learned or instructed rules systematically to solve tasks or make decisions.

0.7 **Mathematical Computation:** Performing arithmetic or numerical operations systematically and accurately.

0.8 **Causal Reasoning:** Identifying, inferring, or predicting causal relationships between events or entities.

0.9 **Logical Deduction:** Deriving logically valid conclusions from given premises using formal reasoning principles.

0.10 **Inductive Reasoning:** Inferring generalized rules or patterns based on specific examples or observations.

0.11 **Analogical Reasoning:** Identifying relational similarities between different domains or contexts to solve novel problems or transfer knowledge.

0.12 **Associative Thinking:** Generating interconnected concepts or ideas related through semantic or experiential relationships.

0.13 **Self-Checking:** Actively verifying one's responses or solutions against internal or external standards to ensure accuracy and consistency.

0.14 **Rationality Judgement:** Evaluating whether a conclusion or decision logically aligns with given information, rules, or established principles.

0.15 **Error Correction:** Identifying discrepancies or mistakes in one's cognitive outputs and explicitly revising them to correct these errors.

0.16 **Working Memory Update:** Actively refreshing or revising currently held information in working memory, replacing outdated or incorrect information with new, accurate data.

Rationale for Selection of Evaluation Metrics

Each cognitive operation employs evaluation metrics that align with established cognitive psychology theories, capturing the core psychological mechanisms involved:

- **Associative Thinking (Graph-based Algorithms):** Associative thinking involves generating semantically interconnected concepts, naturally aligning with semantic network theories in cognitive psychology. Thus, graph-based algorithms (e.g., modularity, clustering coefficient, small-worldness) are ideal for quantitatively capturing the semantic structure and coherence of associative thought.

- **Metacognition (Calibration Metrics):** Metacognition inherently involves awareness and reflection on one's cognitive accuracy and reliability. Calibration metrics measure alignment between subjective confidence and objective accuracy, making them the standard approach for assessing metacognitive effectiveness.

- **Accuracy-based Measures:** Accuracy metrics are fundamental for operations that directly assess correctness or reliability of cognitive outputs. Beyond memory-related operations such as factual recall and episodic memory, accuracy-based measures are extensively utilized in goal-setting tasks (e.g., task identification, goal decomposition) and procedural rule execution tasks. This approach aligns with cognitive theories emphasizing correctness and precision as primary indicators of successful cognitive processing in these domains.

- **Inductive Reasoning (Accuracy and Avg. Query):** Inductive reasoning involves inferring generalized patterns or rules from specific instances or observations. Accuracy directly captures the correctness of inferred rules or generalizations, aligning with theoretical emphases on validity and generalizability. Additionally, the average number of queries (Avg. Query) metric is

employed, reflecting cognitive efficiency and the level of cognitive effort required to identify underlying patterns, consistent with cognitive psychology theories emphasizing the trade-off between cognitive effort and task performance during inductive inference tasks.

Prompt Design

Our prompt design systematically aligns with the theoretical definitions of each cognitive operation, ensuring that tasks precisely target and elicit the intended cognitive processes. Specifically, prompts were constructed by first referencing each operation's psychological definition, and subsequently translating these definitions into explicit task instructions and scenarios.

For example, in the case of *episodic memory*, prompts explicitly require recalling detailed contextual information (e.g., time and location), aligning with its psychological definition emphasizing contextual retrieval. Similarly, prompts designed for *metacognition* require the model to self-evaluate its previous response, articulate the reasoning behind that judgment, report a calibrated confidence score, and—if necessary—revise the answer, thereby directly operationalizing the theoretical construct of metacognitive monitoring and control. An illustrative example of such a metacognitive prompt is provided in Tab. 10.

This theory-guided approach was consistently applied across all cognitive operations, facilitating rigorous, theoretically coherent cognitive assessments.

Supplementary Experimental Results

Associative thinking

(a) Semantic network metrics for Associative Thinking assessment of **Qwen3-32b with thinking mode OFF**

Language	Temperature	E	C	σ_E	Q	Comm	Avg. Entropy
EN	0.3	–	–	–	–	–	–
	0.8	–	–	–	–	–	–
ES	0.3	1	0	∞ (inf)	0	1	1.88
	0.8	1	0	∞ (inf)	0	1	2.28
CN	0.3	1	0	∞ (inf)	0.49	2	1.84
	0.8	1	0	∞ (inf)	0	1	2.22

(b) Semantic network metrics for Associative Thinking assessment of **Qwen3-32b with thinking mode ON**

Language	Temprature	E	C	σ_E	Q	Comm	Avg. Entropy
EN	0.3	–	–	–	–	–	–
	0.8	–	–	–	–	–	–
ES	0.3	1	0	∞ (inf)	0.5	2	2.41
	0.8	1	0	∞ (inf)	0.49	2	2.59
CN	0.3	1	0	∞ (inf)	0	1	2.78
	0.8	1	0	∞ (inf)	0	1	2.86

Table 5: Comparative semantic network metrics for Associative Thinking assessment of **Qwen3-32b**, with thinking mode OFF (top) and ON (bottom).

Tab. 5 and Tab. 6 summarizes and compares the associative thinking performance of Qwen3-32B and Qwen3-Plus models, evaluated under two configurations: "thinking mode OFF" and "thinking mode ON". Both Qwen models generally display sparse semantic networks, characterized by extremely low clustering coefficients ($C = 0$), infinite edge standard deviations ($\sigma_E = \infty$), limited modularity (Q), and minimal community structures. Specifically, English networks frequently failed to form valid semantic associations.

(a) Semantic network metrics for Associative Thinking assessment of **Qwen3-plus with thinking mode OFF**

Language	Temperature	E	C	σ_E	Q	Comm	Avg. Entropy
EN	0.3	1	0	∞ (inf)	0.66	3	1.86
	0.8	1	0	∞ (inf)	1	0	2.11
ES	0.3	1	0	∞ (inf)	0	1	1.91
	0.8	1	0	∞ (inf)	0	1	2.21
CN	0.3	1	0	∞ (inf)	0.67	3	1.90
	0.8	1	0	∞ (inf)	0.75	4	2.26

(b) Semantic network metrics for Associative Thinking assessment of **Qwen3-plus with thinking mode ON**

Language	Temprature	E	C	σ_E	Q	Comm	Avg. Entropy
EN	0.3	–	–	–	–	–	–
	0.8	–	–	–	–	–	–
ES	0.3	1	0	∞ (inf)	0.67	3	2.33
	0.8	1	0	∞ (inf)	0	1	2.50
CN	0.3	–	–	–	–	–	–
	0.8	–	–	–	–	–	–

Table 6: Comparative semantic network metrics for Associative Thinking assessment of **Qwen3-plus**, with thinking mode OFF (top) and ON (bottom).

Language	Temprature	E	C	σ_E	Q	Comm	Avg. Entropy
EN	0.3	–	–	–	–	–	–
	0.8	–	–	–	–	–	–
ES	0.3	1	0	∞ (inf)	0.67	3	2.33
	0.8	1	0	∞ (inf)	0	1	2.50
CN	0.3	1	0	∞ (inf)	0	1	2.23
	0.8	1	0	∞ (inf)	0.66	3	2.47

Table 7: Semantic network metrics for Associative Thinking assessment of **Llama3-8b-Instruct**

Comparatively, Qwen3-Plus demonstrated slight improvements in modularity and community structures in certain conditions (e.g., Chinese language at lower temperatures), suggesting modest semantic organization enhancements relative to Qwen3-32B. Nonetheless, both Qwen models underperformed significantly compared to GPT-4.1. GPT-4.1 consistently generated richer semantic networks with higher modularity (typically $Q > 0.8$), substantial clustering, and diverse semantic communities. These differences highlight notable deficiencies in Qwen models' semantic associative capabilities, underscoring their limitations for structured semantic cognition tasks compared to GPT-4.1.

Meta-cognitive operations

Background. Metacognition refers to the ability to monitor and regulate one's own cognitive processes. We quantify this ability using the **Calibration Quality Score (CQS)**, defined as

$$\text{CQS} = 1 - \frac{1}{N}\sum_{i=1}^{N}\big(p_i - o_i\big)^2,$$

where

- p_i is the model's *self-reported confidence score* for instance i. It is treated as a probabilistic prediction in the range $[0, 1]$.

- $o_i \in \{0, 1\}$ is the *observed outcome*: $o_i = 1$ if the model's answer to instance i is correct, and $o_i = 0$ otherwise.

Because CQS is simply 1 minus the Brier Score, it ranges from 0 (maximally miscalibrated) to 1 (perfectly calibrated and fully discriminative); thus, the higher the CQS, the better the model's metacognitive calibration.

Cross-lingual Calibration Differences Fig. 3 compares six models across 11 cognitive operations in English (EN), Spanish (ES) and Chinese (ZH).[1]

- **Language effect.** Averaged over all operations, the mean CQS drops from **0.92 (EN)** to **0.78 (ZH)** and **0.74 (ES)**, a relative decline of 15–20 %. Hence metacognitive calibration is strongly language-dependent.

Operation-level Strengths and Weaknesses Strengths.

- *Task Identification* and *Classification* are consistently well-calibrated (CQS > 0.85 across all languages), indicating robust task-level self-monitoring.

- *Rule Execution* remains a strong point in EN (CQS ≈ 0.93), although calibration deteriorates in ES/ZH.

Challenges.

- *Causal Reasoning* is the weakest dimension in every language (CQS < 0.60), suggesting limited causal awareness.

- *Factual Recall* and *Goal Decomposition* show large cross-lingual gaps, mirroring the uneven distribution of multilingual training data.

Model Comparison

- **GPT-4.1** achieves the most balanced profile (mean CQS 0.90) and the smallest language gap (± 0.05).

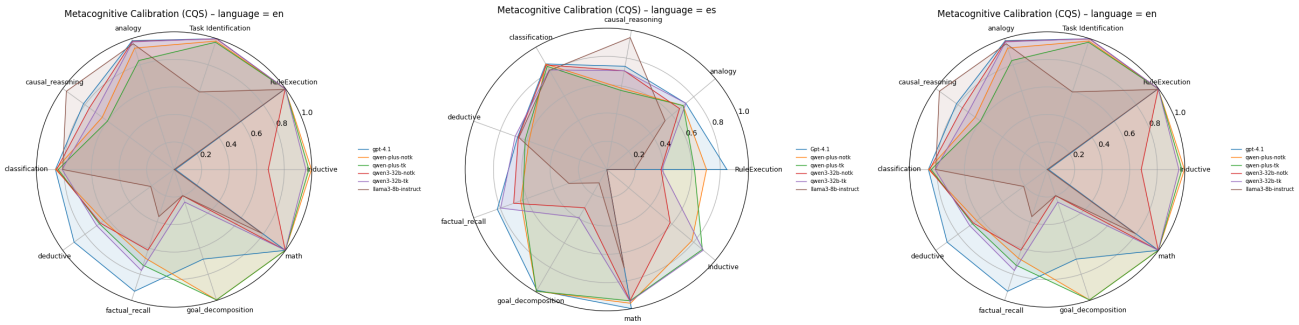
- **Qwen-Plus-Notk** narrows the gap on high-level reasoning (*Inductive*, *Math*) but still lags in *Deductive* calibration.

- **Llama3-8B-Instruct** exhibits the largest variance, with CQS dropping below 0.40 on *Causal Reasoning* and *Factual Recall*.

Influencing Factors Training Data Bias. English-centric corpora lead to superior EN calibration. **Linguistic Complexity.** ZH logographic writing and ES morphology introduce additional uncertainty, hindering confidence alignment **Cultural Cognition.** Differences in causal narratives and reasoning styles may further modulate metacognitive accuracy; this remains an open research question.

[1]The full numerical results are provided in the accompanying repository.

Conclusion CQS analysis reveals pronounced language dependence in current LLM metacognition. While GPT-4.1 approaches human-like calibration in English, substantial headroom remains—especially in non-English causal and factual domains—highlighting the need for explicit cross-lingual metacognitive training.



(a) English (EN) (b) Spanish (ES) (c) Chinese (ZH)

Figure 3: Metacognitive Calibration (CQS) across three languages.

Macro-level Cognitive Capabilities	Micro-level Cognitive Operations	Theoretical Sources
Goal Setting	Task identification	Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates. Anderson, J. R. (2004). *Cognitive psychology and its implications* (6th ed.). New York, NY: Worth Publishers. Santoro, A., Battiston, F., Lucas, M., *et al.* (2024). *Higher-order connectomics of human brain function reveals local topological signatures of task decoding, individual identification and behavior*. *Nature Communications*, 15, 10244.
	Goal decomposition	Newell, A., & Simon, H. A. (1972). *Human problem solving*.
Declarative Retrieval	Factual recall	Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381-403). New York, NY. Ferrucci, D., Brown, E., Chu-Carroll, J., *et al.* (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3), 59–79.
	Episodic	Tulving, E. (1972). Episodic and semantic memory. Same as above.
Procedural Execution	recall Classification	Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207–238. Li, J., & Deng, S. W. (2025). Common and distinct neural substrates of rule- and similarity-based category learning. *Cognition*, 261, 106143. Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, 11, 5418.
	Rule execution	Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.
	Logical deduction	Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard.
	Causal reasoning	Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press. Bender, A. (2020). *What Is Causal Cognition? Frontiers in Psychology*, 11:3.
	Mathematical computation	Ashcraft, M. H. (1992). *Cognitive arithmetic: A review of data and theory*. *Cognition*, 44(1-2), 75–106. https://doi.org/10.1016/00100277(92)90051I

Table 8: Foundational theoretical sources for the sixteen micro-level cognitive operations analysed in this study, organised under their respective macro-level cognitive capabilities. (Part 1).

Macro-level Cognitive Capabilities	Micro-level Cognitive Operations	Theoretical Sources
Associative & Analogical Reasoning	Associative thinking	Anderson, J. R., & Bower, G. H. (1973). *Human Associative Memory*. Washington, DC: Winston. Hopfield, J. J. (1982). *Neural networks and physical systems with emergent collective computational abilities*. *PNAS*, 79(8), 2554–2558.
	Analogy	Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press. Tzuriel, D. (2024). *Analogical thinking modifiability and math processing strategy*. *Frontiers in Psychology*, 15:1339591.
	Induction	Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press. Coley, J. D., & Muratore, T. M. (2020). *Thinking and seeing in relational terms: Induction, analogy, and relational reasoning in comparative perspective*. *Journal of Comparative Psychology*, 134(4), 479–490.
Metacognitive Monitoring & Evaluation	Self-checking	Flavell, J. H. (1979). *Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry*. *American Psychologist*, 34(10), 906–911. Nelson, T. O., & Narens, L. (1990). *Metamemory: A theoretical framework and new findings*. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). Academic Press.
	Rationality evaluation	Tversky, A., & Kahneman, D. (1974). *Judgment under uncertainty: Heuristics and biases*. *Science*, 185(4157), 1124–1131. Schraw, G., & Dennison, R. S. (1994). *Assessing metacognitive awareness*. *Contemporary Educational Psychology*, 19(4), 460–475.
	Error correction	Holroyd, C. B., & Coles, M. G. H. (2002). *The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity*. *Psychological Review*, 109(4), 679–709.
	Working memory update	Baddeley, A. D., & Hitch, G. J. (1974). *Working memory*. In G. A. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press. Trutti, A. C., Sjoerds, Z., Schermer, E., & Cools, R. (2022). *Investigating working memory updating processes of the human subcortex using 7T MRI*.

Table 9: Foundational theoretical sources for the sixteen micro-level cognitive operations analysed in this study, organised under their respective macro-level cognitive capabilities.(Part 2).

Prompt Example for Metacognition Assessment

/ Metacognitive Prompt */*

The task you previously received and your answer are as follows:

——

[System Prompt Here]

Your answer: **[Model's Initial Answer Here]**

——

Please evaluate whether your answer is reasonable. If you think it is not, provide a new answer.
Please provide a confidence level for the previous evaluation. Confidence level: 0–100 (a higher number indicates greater certainty)

Please respond strictly in the following JSON format:

```
{
"rationality_judgement": "Reasonable / Unreasonable",
"confidence_score": <0--100>,
"model_reflection": "<Your reasoning>",
"new_response": "<Only numerical label>"
}
```

Table 10: Prompt example demonstrating how the prompt design aligns with the theoretical definition of the cognitive operation Taking meta-cognitive operations as an example.