System 1.5

Designing Metacognition in AI

System-2 Reasoning at Scale Workshop
@ NeurIPS 2024

👨‍💻👩‍💻 Co-authors (+background)

Nick Oh
socius

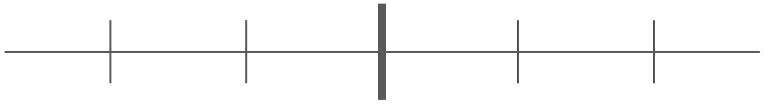
Prof. Fernand Gobet
Centre for Philosophy of Natural and Social Science (CPNSS), LSE

🧠 System 1.5

System 1 | System 2

🧠 System 1.5

System 1.5

System 1  System 2

System 1.5

Three key components:

1. **Monitor** (**M**), assessing "familiarity" score *ƒ*
   - where *ƒ* = *problem-type* and *solution-pattern*
2. **Generator** (**G**), producing potential *k* solutions
   - where *k* = C(1 – *ƒ*) and C = hyperparameter
3. **Evaluator** (**V**), verifying the quality of those solutions
   - Optimisation problem

NOTE: *The paper prioritises* ***theoretical*** *description grounded in psychological and cognitive insights over mathematical formalisations (e.g., Direct Preference Optimisation)*

Initialization: Let $\mathcal{M}_{\text{BASE}}$ be a pretrained base model and $\mathcal{D}_{\text{SFT}} = \{(x_i, y_i)\}_{i=1}^{N}$ be an initial dataset where x_i represents problem description and y_i represents corresponding solution. We obtain:

$$\mathcal{M}_{\text{SFT}} = \text{finetune}(\mathcal{M}_{\text{BASE}}, \mathcal{D}_{\text{SFT}})$$

Iterative Process (for iterations $t = 1, \ldots, T$):

1. For each problem x_i, generate k candidate solutions (Cobbe et al., 2022):

$$\{\hat{y}_{ij} \sim \mathcal{M}_t(y|x_i)\}_{j=1}^{k}$$

2. Construct three datasets:

$$\mathcal{D}_{\text{GENERATE}_t} = \{(x_i, \hat{y}_i) | z_{ij} = \text{preferred}\}$$

where z_{ij} indicates preference

$$\mathcal{D}_{\text{EVALUATE}_t} = \{(x_i, \hat{y}_{ij}, z_{ij})\}$$

containing all solutions with preferences

$$\mathcal{D}_{\text{MONITOR}_t} = \{(x_i, \hat{y}_{ij}, f_{ij})\}$$

where f_{ij} captures "familiarity".

3. Update model:

$$\mathcal{M}_t = \text{finetune}(\mathcal{M}_{\text{BASE}}, \mathcal{D}_{\text{GENERATE}_{t-1}})$$

At final iteration T, we obtain three specialised functions:

$$\begin{aligned} \mathcal{G}_T &= \text{train}(\mathcal{D}_{\text{GENERATE}_{T-1}}) && \text{(Generator)} \\ \mathcal{V}_T &= \text{train}(\mathcal{D}_{\text{EVALUATE}_{T-1}}, \text{preference optimisation}) && \text{(Evaluator)} \\ \mathcal{M}_T &= \text{train}(\mathcal{D}_{\text{MONITOR}_{T-1}}, \text{familiarity assessment}) && \text{(Monitor)} \end{aligned}$$

Inference (for input x): At inference time, we adapt our strategy based on how familiar the problem is:

1. Compute familiarity: $f = \mathcal{M}_T(x)$
2. Determine strategy based on familiarity:

$$\text{out}(x) = \begin{cases} \mathcal{G}_T(x) & \text{if } f > \theta_h \\ \text{argmax}_{y \in \{\mathcal{G}_T(x)_j\}_{j=1}^{k}} \mathcal{V}_T(x, y) & \text{if } \theta_l < f \leq \theta_h \\ \text{argmax}_{y \in \mathcal{Y}} \mathcal{V}_T(x, y) & \text{if } f \leq \theta_l \end{cases}$$

where $k = \lceil C(1 - f) \rceil$ and $\mathcal{Y} = \{\mathcal{G}_T(x)_j\}_{j=1}^{k_{\max}} \cup \{\mathcal{G}_T(x|h) : h \in \text{System-2}(x)\}$

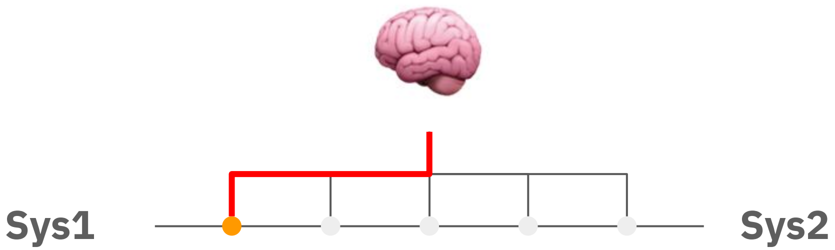
Chess, **as an Intuitive Explanation**

1. Compute familiarity: $f = \mathcal{M}_T(x)$
2. Determine strategy based on familiarity:

$$\text{out}(x) = \begin{cases} \mathcal{G}_T(x) & \text{if } f > \theta_h \\ \text{argmax}_{y \in \{\mathcal{G}_T(x)_j\}_{j=1}^{k}} \mathcal{V}_T(x, y) & \text{if } \theta_l < f \leq \theta_h \\ \text{argmax}_{y \in \mathcal{Y}} \mathcal{V}_T(x, y) & \text{if } f \leq \theta_l \end{cases}$$

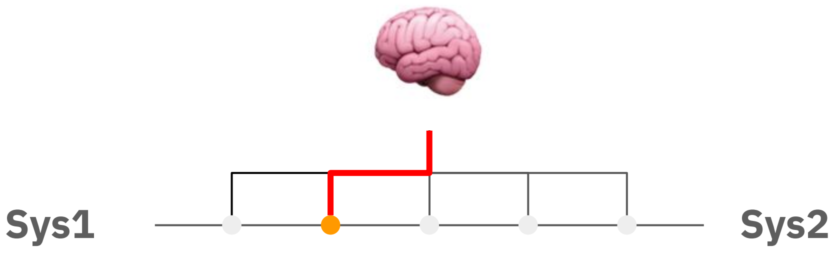
where $k = \lceil C(1 - f) \rceil$ and $\mathcal{Y} = \{\mathcal{G}_T(x)_j\}_{j=1}^{k_{\max}} \cup \{\mathcal{G}_T(x|h) : h \in \text{System-2}(x)\}$

♟️ Chess, **as an Intuitive Explanation**



$$f > \vartheta_h$$

♟ Chess, **as an Intuitive Explanation**

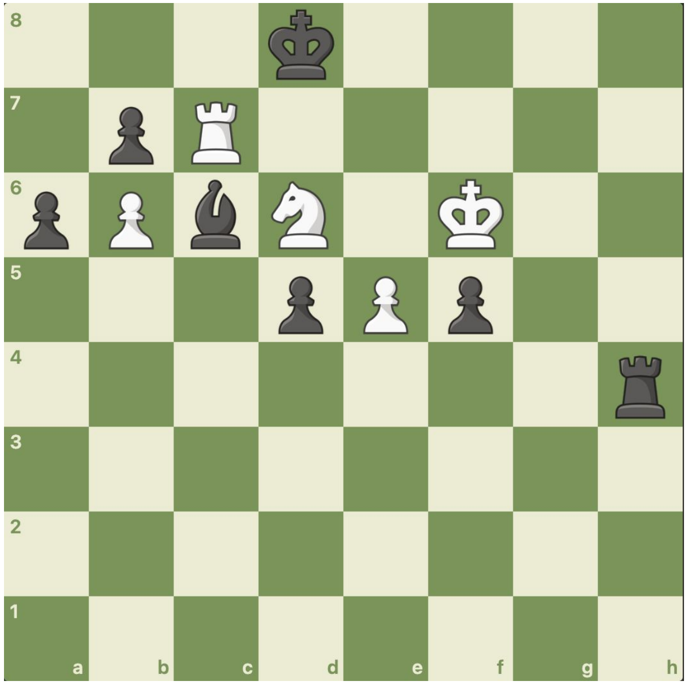


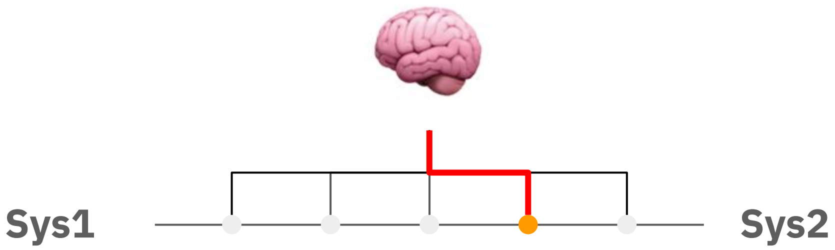
$$\vartheta_l < f \leq \vartheta_h$$

e.g., $k = C(1 - f)$, where $f = 0.8$ and $C = 10$

$$\mathrm{argmax}_{y \in \{\mathcal{G}_T(x)_j\}_{j=1}^{k}} \mathcal{V}_T(x, y)$$

♟️ Chess, **as an Intuitive Explanation**





$$f \leq \vartheta_l$$

$$\mathrm{argmax}_{y \in \mathcal{Y}} \mathcal{V}_T(x, y)$$

$$\mathcal{Y} = \{\mathcal{G}_T(x)_j\}_{j=1}^{k_{\max}} \cup \{\mathcal{G}_T(x|h) : h \in \text{System-2}(x)\}$$

💡 Contributions

(the first half) Finding theoretical consensus on System-1

(the second half) Revisiting cognitive science theories and derive architectural principles for AI systems.

🛠️ Limitations and Future Work

Non-technical nature

“familiarty” formalisation

System-2 discussion

Thank You!

nick.sh.oh@socius.org