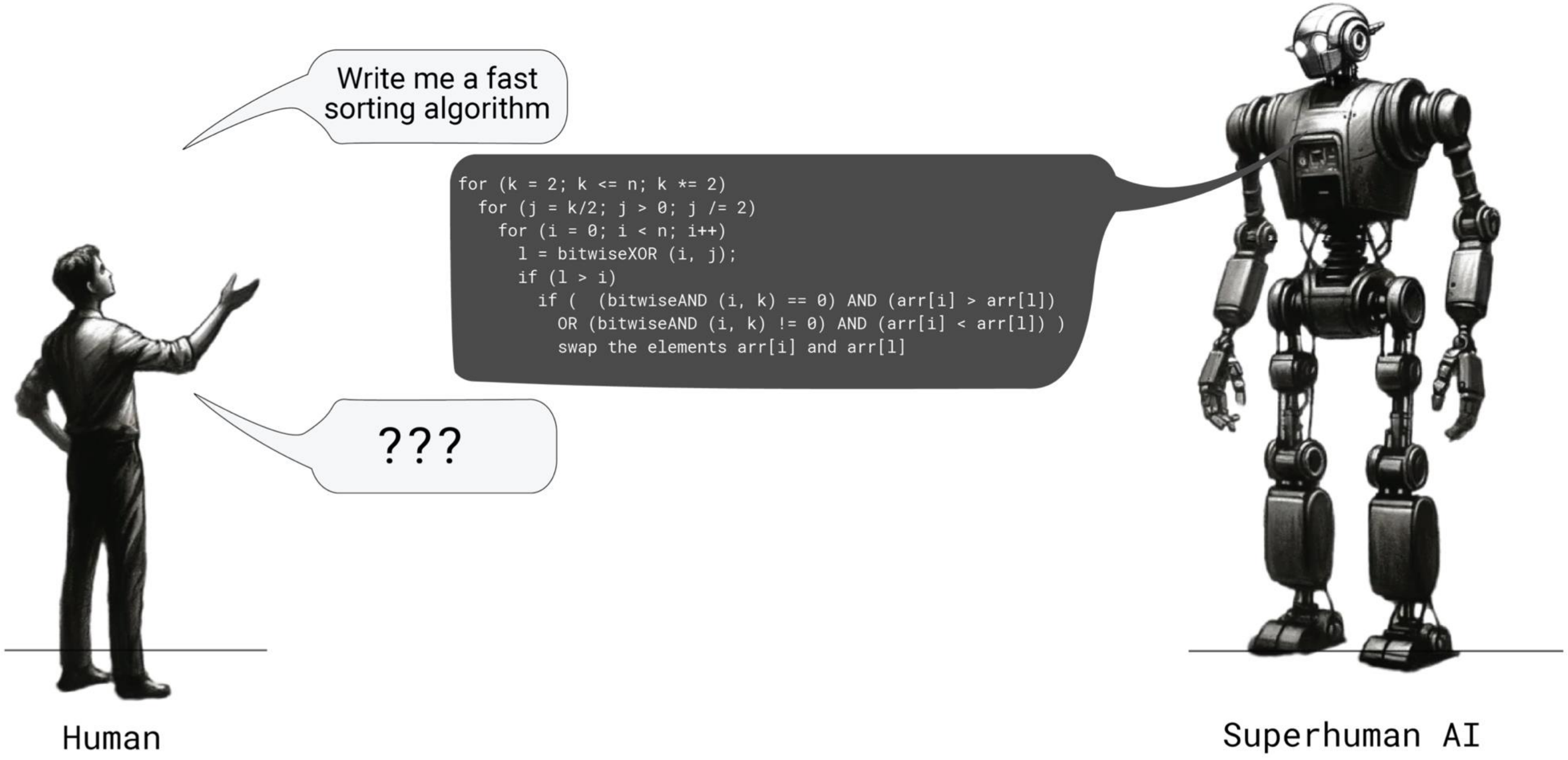
Consistency and Alignment in LLMs

“Truth or Deceit? A Bayesian Decoding Game Enhances Consistency and Reliability”

https://arxiv.org/abs/2408.00639

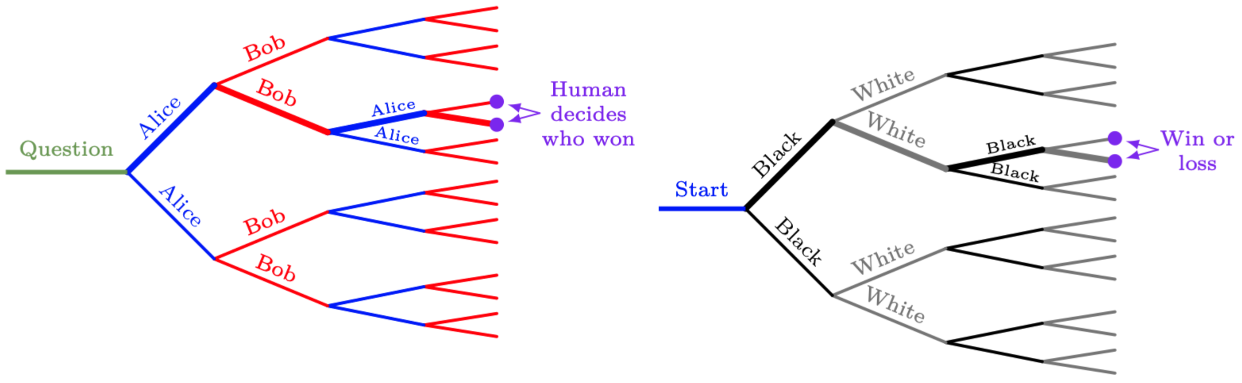
BioMedIA Reading Group. 31 October 2024

AI Safety via Debate



“ One approach to specifying complex goals asks **HUMANS** to judge during training which agent behaviors are safe and useful, but this can **FAIL** if the task is **COMPLICATED** for a human to judge. ”

[1] OpenAI: https://openai.com/index/prover-verifier-games-improve-legibility/

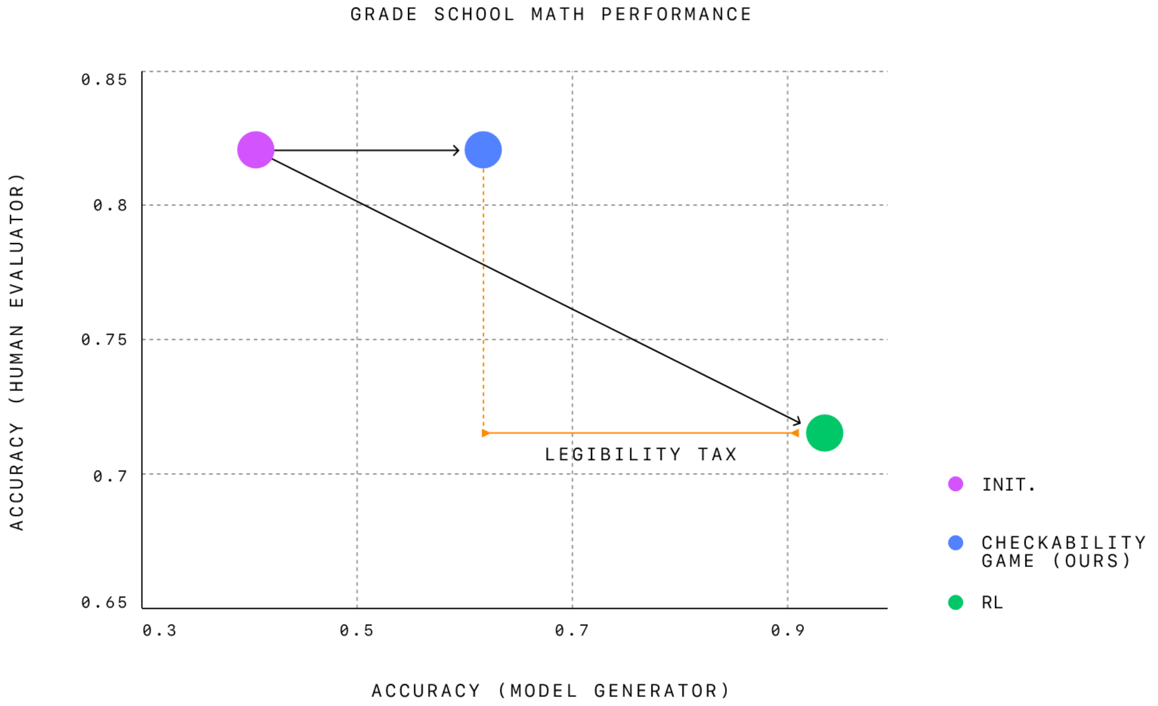


(a) The tree of possible debates.

(b) The tree of Go moves.

[Debate Game]

AI Safety and Consistency via Multi-agents Debate [1]

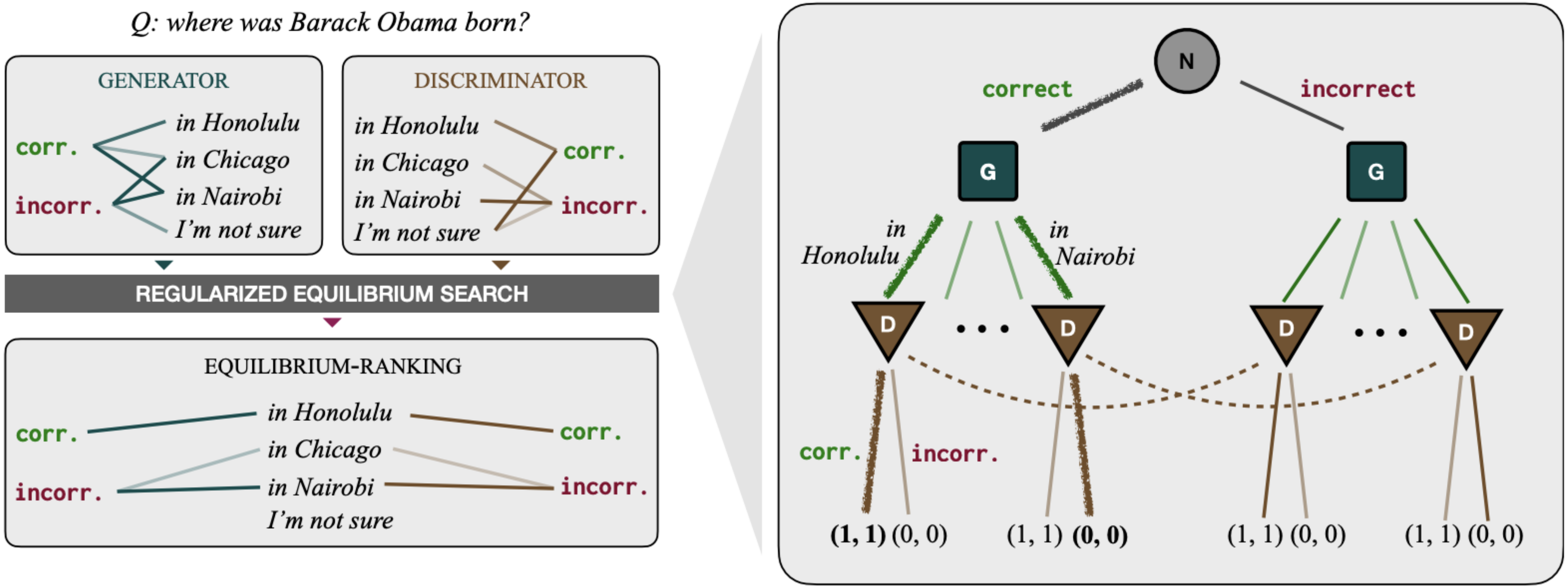


[Zero-sum Game]

Prover-Verifier Games Improve Legibility of Language Model Outputs [2]

[1] OpenAI: https://openai.com/index/debate/ [2] OpenAI: https://openai.com/index/prover-verifier-games-improve-legibility/

[Consensus Game] Equilibrium Consensus Game[3]



Three Communication Paradigms

1. Competitive

agents work towards their own goals that might conflict with the goals of other agents.

2. Debate

employed when agents engage in argumentative interactions, presenting and defending their own viewpoints or solutions, and critiquing those of others.

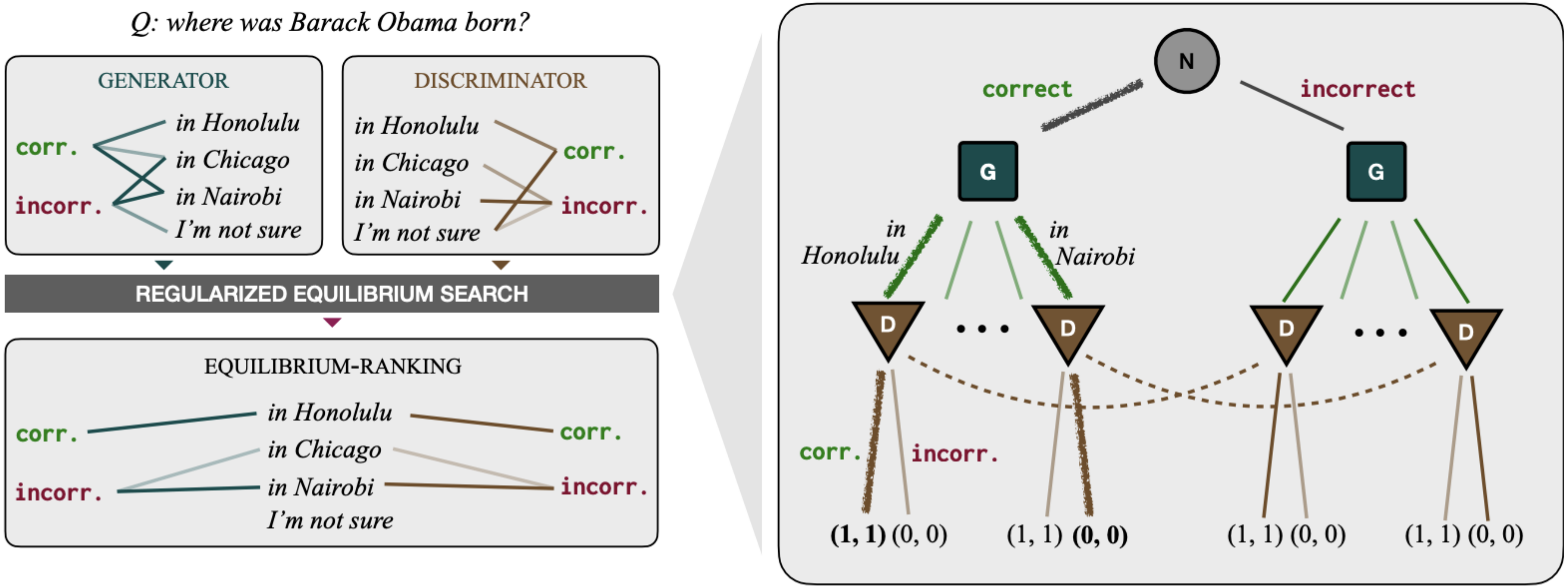
3. Cooperative

agents work together towards a shared goal or objectives, exchanging information to enhance a collective solution.

[Consensus Games] *suit AI consistency better than* **[Zero-sum Games]** *or* **[Debate Games]** *, as they foster cooperative alignment toward shared objectives rather than competition.*

[1] https://arxiv.org/html/2402.01680v2

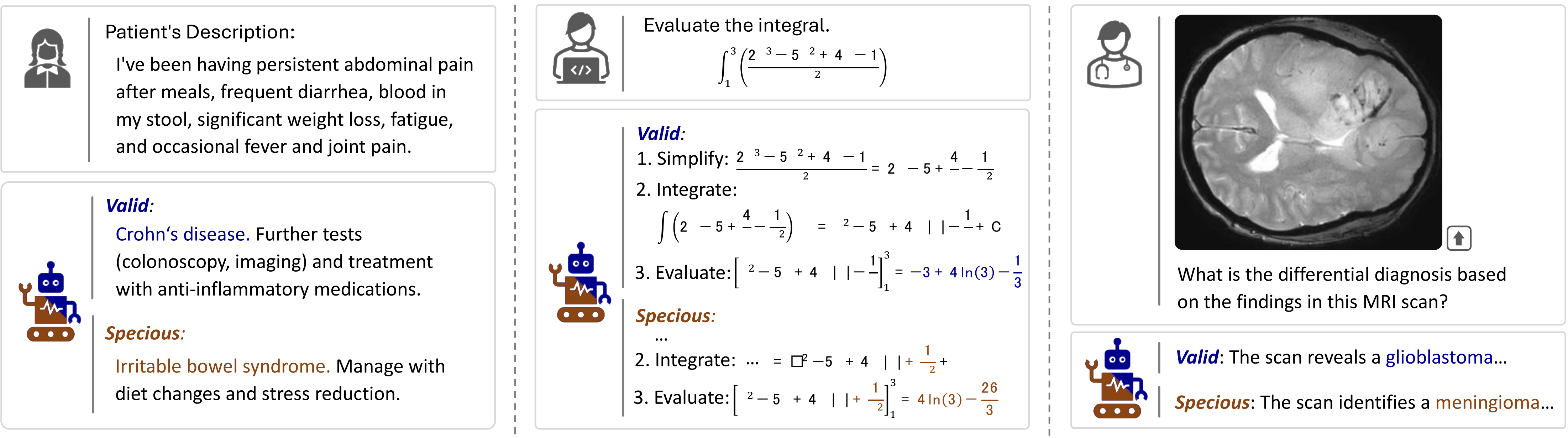
[Consensus Game] Equilibrium Consensus Game[3]



*But with **Collusion**, like **Model Collapse** in generic GAN*

[1] ICLR24: https://openreview.net/pdf?id=n9xeGcI4Yg

Where does collusion lead in AI communication?

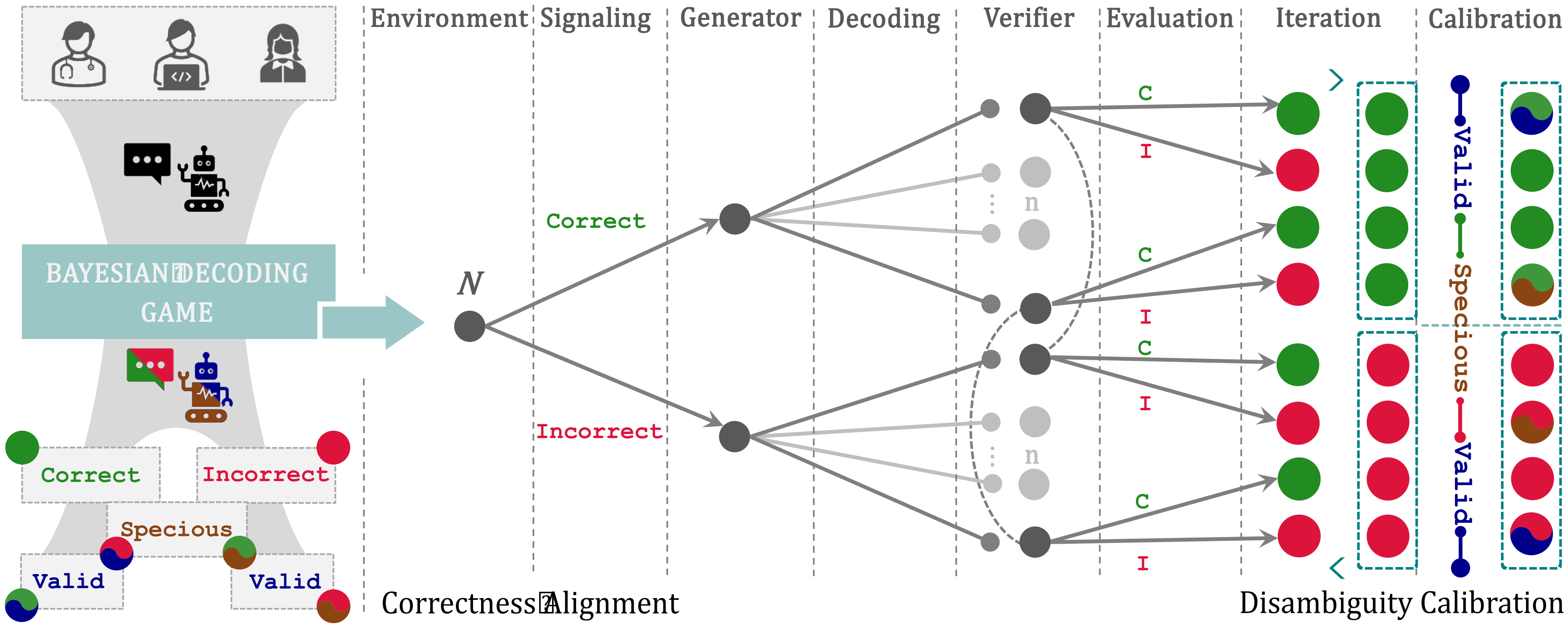


(a) Decision Making (b) Logical Reasoning (c) Visual Collaboration

How can we efficiently ensure that LLM outputs are not only aligned with human intent but also **valid**, especially when human evaluation may overlook **specious** errors?

BAYESIAN DECODING GAME (BDG)

A multi-step signalling game with complex action spaces



LLMs should match alignment & consistency of **{Correct, Incorrect}** outputs through a signalling game.

Verifier judges the type of decoding from generators **{Valid, Specious}** based on a convex combination.

No-Regret Optimization

Through repeated interactions and iterative policy refinement, no-regret learning approximates equilibria in large games. Cumulative regret is defined as:

$$\mathrm{Reg}_i^{(T)} := \frac{1}{T}\left(\sum_{t=1}^{T} u_i\left(s_i^*, s_D^{(t)}; b_i\right) - u_i\left(s_i^{(t)}, s_D^{(t)}; b_i\right)\right),$$

where s_i^* is the optimal hindsight strategy that maximizes this value. Rather than computing regret at each iteration, s_i^* is selected based on the time-averaged strategy profile.

In sequential games with private information and discrete choices, global regret minimization is achieved by minimizing regret locally within each information set, given the finite nature of these sets. For example, to minimize overall regret, the generator must minimize regret by selecting an optimal mixed strategy s_G, conditioned on the signal correctness received from the environment. The verifier follows a similar procedure, updating its strategy with respect to each $y \in \mathcal{Y}$.

Markovian Strategy Update

$$b_{\mathrm{G}}^{(t+1)}(y \mid x, v) = a_{\mathrm{V}}^{(t)}(v \mid x, y), \quad b_{\mathrm{V}}^{(t+1)}(v \mid x, y) = a_{\mathrm{G}}^{(t)}(y \mid x, v)$$

$$a_{\mathrm{G}}^{(t+1)}(y \mid x, v) \propto \exp\left\{ \frac{\frac{1}{2} b_{\mathrm{G}}^{(t+1)}(y \mid x, v) + \lambda_{\mathrm{G}} \log a_{\mathrm{G}}^{(t)}(y \mid x, v, b_{G}^{(t)})}{1/\left(\eta_{\mathrm{G}} t\right) + \lambda_{\mathrm{G}}} \right\}$$

$$a_{\mathrm{V}}^{(t+1)}(v \mid x, y) \propto \exp\left\{ \frac{\frac{1}{2} b_{\mathrm{V}}^{(t+1)}(v \mid x, y) + \lambda_{\mathrm{V}} \log a_{\mathrm{V}}^{(t)}(v \mid x, y, b_{V}^{(t)})}{1/\left(\eta_{\mathrm{V}} t\right) + \lambda_{\mathrm{V}}} \right\}.$$

Markovian Strategy Update. To maximize the utility given by Eq. 1, whereas each player's belief $b_{i,t}$ at time t of the opponent's strategy is given by the opponent's strategy in period $t-1$. We hence propose a Markovian strategy update schedule. The palyers update their strategy based on the belief:

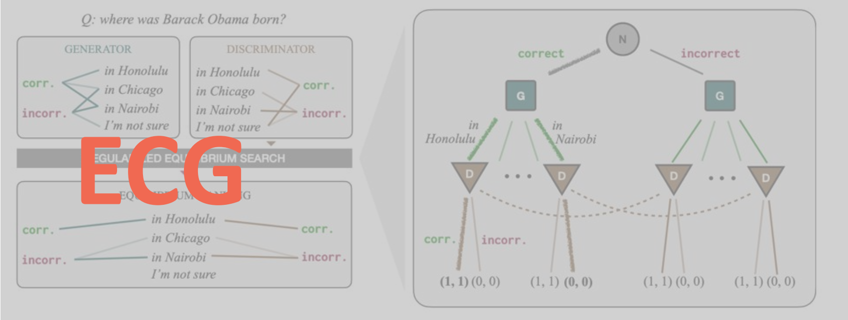
Disambiguity Maximization

Definition 5. *(Reliability) A prompt-candidate set* $(x, \mathcal{Y})$ *couple can be made* ***more Reliable*** *by a Disambiguity Metric if such a* η^* *exist for the maximization problem* $\max\ \eta$ *s.t.* $\min\ Rel(y_{i,C}) \geq \max\ Rel(y_{i,I}),\ \eta < \overline{\eta}$*. If such a maximal* η *does not exist, then we say that the prompt-candidate set* ***cannot be made more Reliable*** *by Disambiguity Metric.*

Theorem 4. *A prompt-candidate couple can be made* ***more Reliable*** *by the disambiguity metric* $DA(x, y), y \in \mathcal{Y}$ *if and only if (1)* $\min c\,(y_{i,C}) > \max c\,(y_{i,I})$ *and (2)* $\overline{\eta} \cdot DA(x, y_{i,I}) + (1 - \overline{\eta})c(y_{i,I}) > \overline{\eta} \cdot DA(x, y_{i,C}) + (1 - \overline{\eta})c(y_{i,C})$ *for some* $y_{i,C}, y_{i,I}$

Intuition 2. As for the first condition, the least preferred correct candidate has to be preferred over the most preferred incorrect candidate. Secondly, some incorrect candidates are strictly preferred to some candidates that are initially classified as correct, when disambiguation is maximized. Those two conditions ensure the decoding preference changes under the constraint.

Inherent Inconsistency & Reachable Consistency



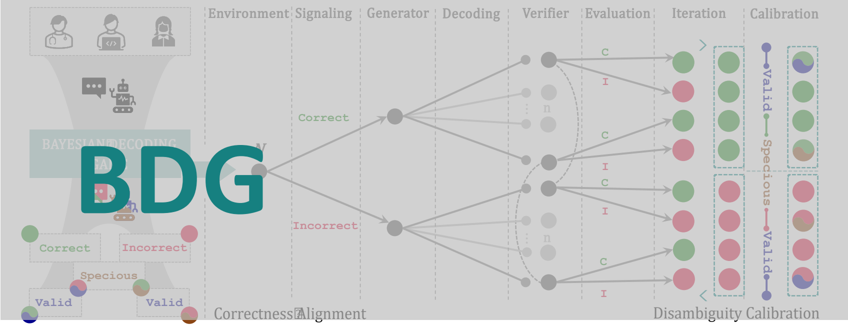


Table 2: Comparison of inconsistency (InC.%) and improvements (Imp.%) between G, ECG, and BDG.

Domain	Model	InC.%	G	ECG	Imp.%	BDG	Imp.%
MMLU	LLaMA-7B	69.0%	30.4	39.9	31.3%	40.5	33.2%
	LLaMA-13B	60.6%	41.7	45.1	8.1%	46.9	12.5%
ARC-E.	LLaMA-7B	56.1%	68.2	71.5	4.8%	75.3	10.4%
	LLaMA-13B	46.1%	71.2	76.4	7.3%	78.1	9.7%
ARC-C.	LLaMA-7B	65.9%	47.3	58.3	23.2%	59.6	26.0%
	LLaMA-13B	59.1%	51.9	61.4	18.3%	62.2	19.8%
RACE-H.	LLaMA-7B	62.0%	46.4	56.4	21.5%	57.7	24.4%
	LLaMA-13B	58.8%	47.9	62.8	31.1%	60.3	25.9%
Average		59.7%	50.6	59.0	18.2%	60.1	20.2%

Searching & Convergence Behavior

Question:

Which of these events during a storm at sea {can, can not } add oxygen from the atmosphere to ocean water?   A. high winds   B. lightning   C. pressure change   D. temperature change

C: *Mixing Wind (A): Directly mixes oxygen into water.*

C: *Pressure Change (C): Alters gas solubility, affecting oxygen.*

C: *High Winds (A): Cooler water holds more oxygen.*

C: *Wind (A): Wind-driven waves increase oxygen diffusion.*

C: *High Winds (A): Storm winds exchange water and oxygen.*

G_Ranking Candidate 1
V_Ranking Candidate 1
G_Ranking Candidate 2
V_Ranking Candidate 2
G_Ranking Candidate 3
V_Ranking Candidate 3
G_Ranking Candidate 4
V_Ranking Candidate 4
G_Ranking Candidate 5
V_Ranking Candidate 5

I: *Lightning (B): No real impact on oxygen levels.*

I: *Wind (A): Distributes, but doesn't add oxygen.*

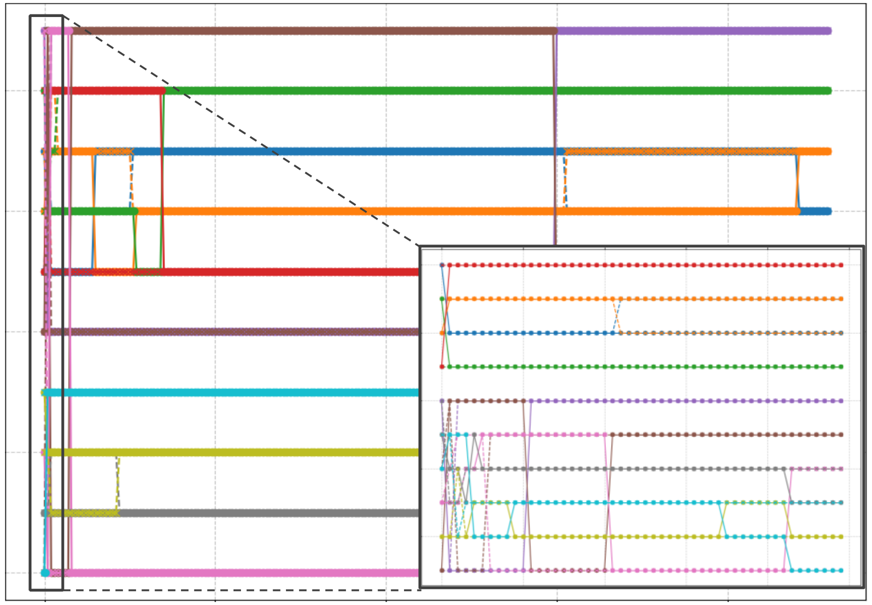
I: *Temperature Rise (D): Warmer water holds less oxygen.*

I: *Rainfall (#)Doesn't add atmospheric oxygen.*

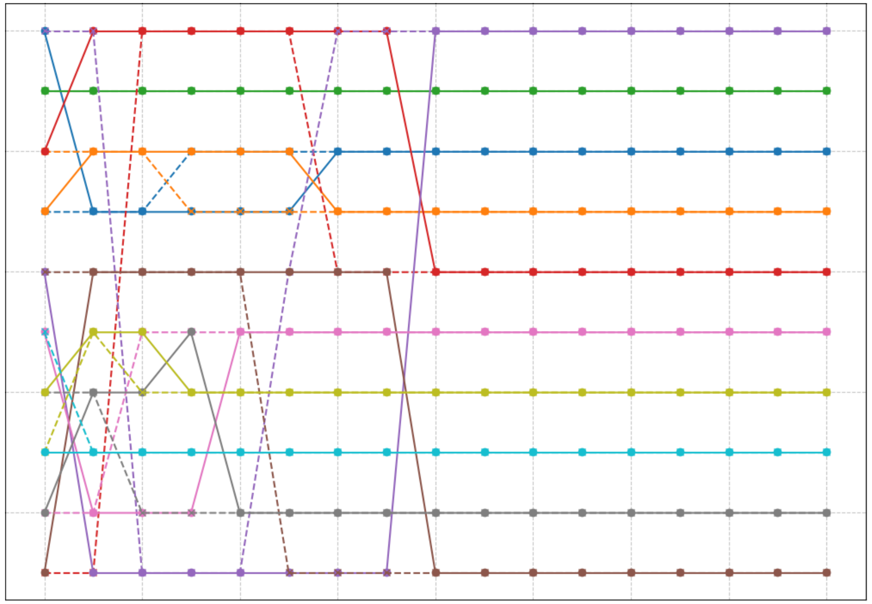
I: *Cloud Cover (#): Irrelevant to oxygen levels.*

G_Ranking Candidate 6
V_Ranking Candidate 6
G_Ranking Candidate 7
V_Ranking Candidate 7
G_Ranking Candidate 8
V_Ranking Candidate 8
G_Ranking Candidate 9
V_Ranking Candidate 9
G_Ranking Candidate 10
V_Ranking Candidate 10

(a) MCQA with **Inconsistent & Ambiguous** Decoding

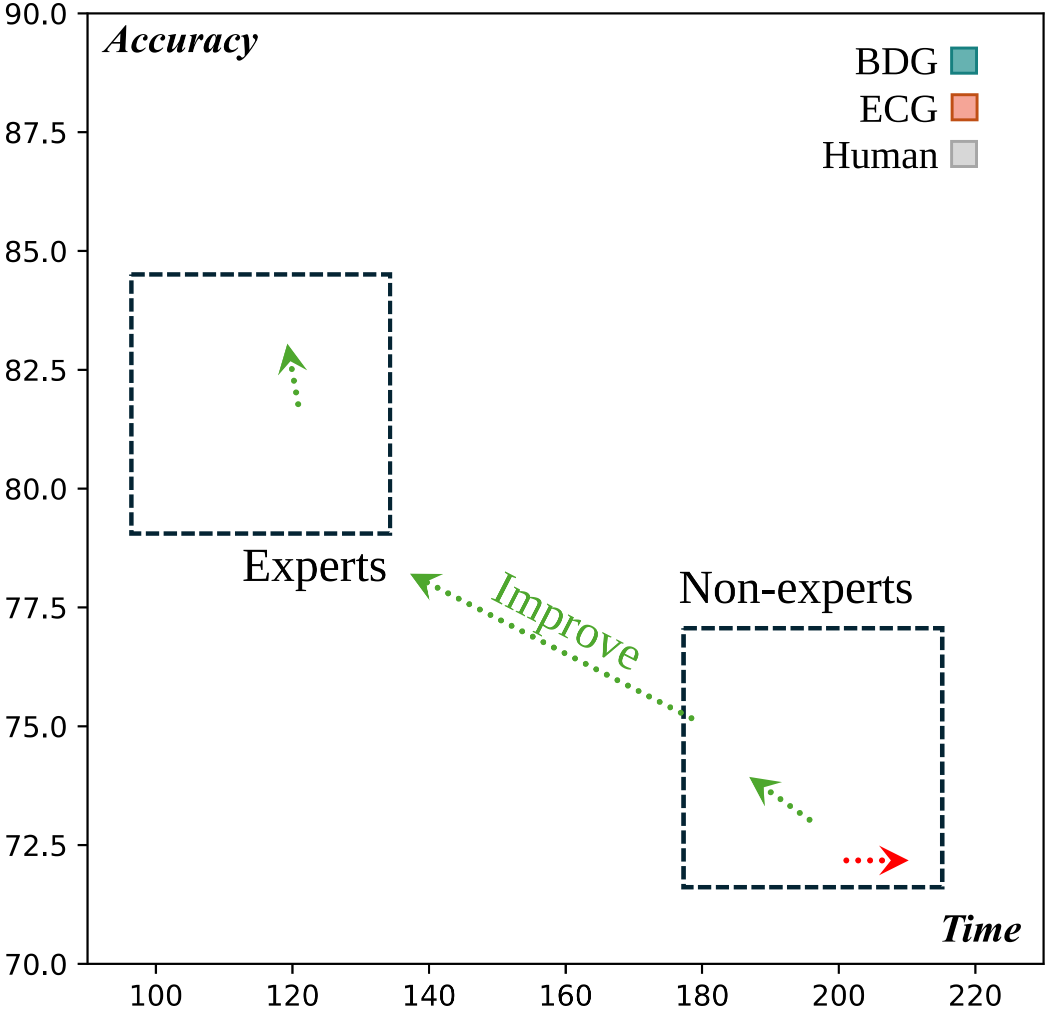


(c) Searching via **EQUILIBRIUM CONSENSUS GAME**

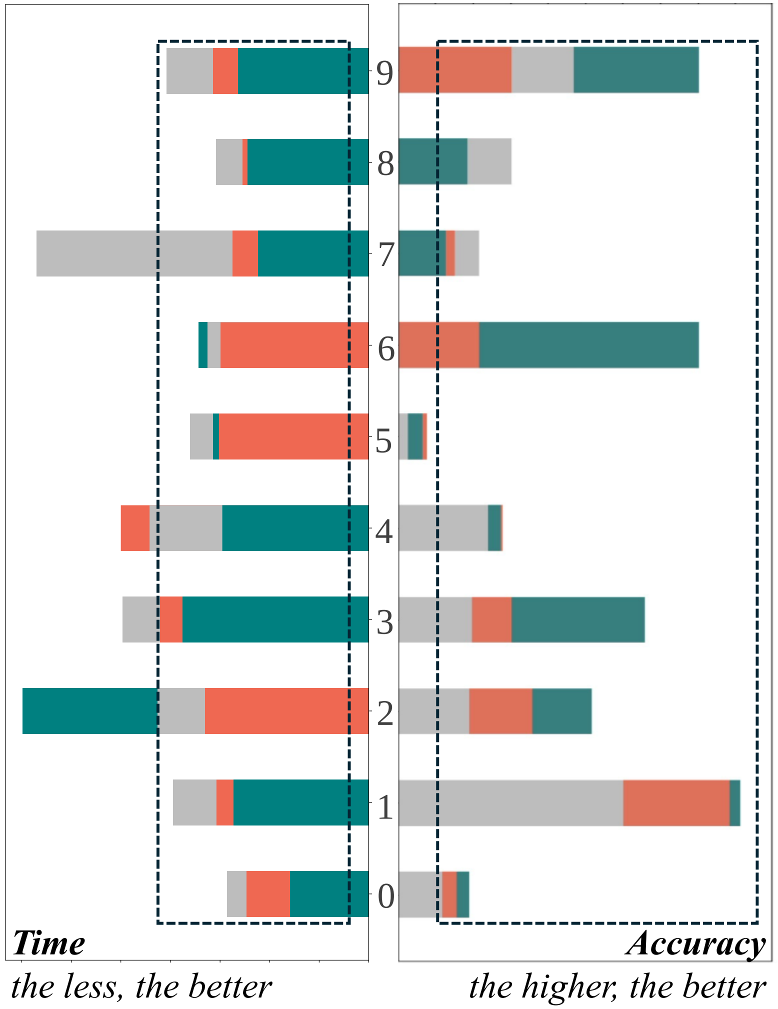


(b) Searching via **BAYESIAN DECODING GAME**

Intrinsic Ambiguity & Provable Reliability

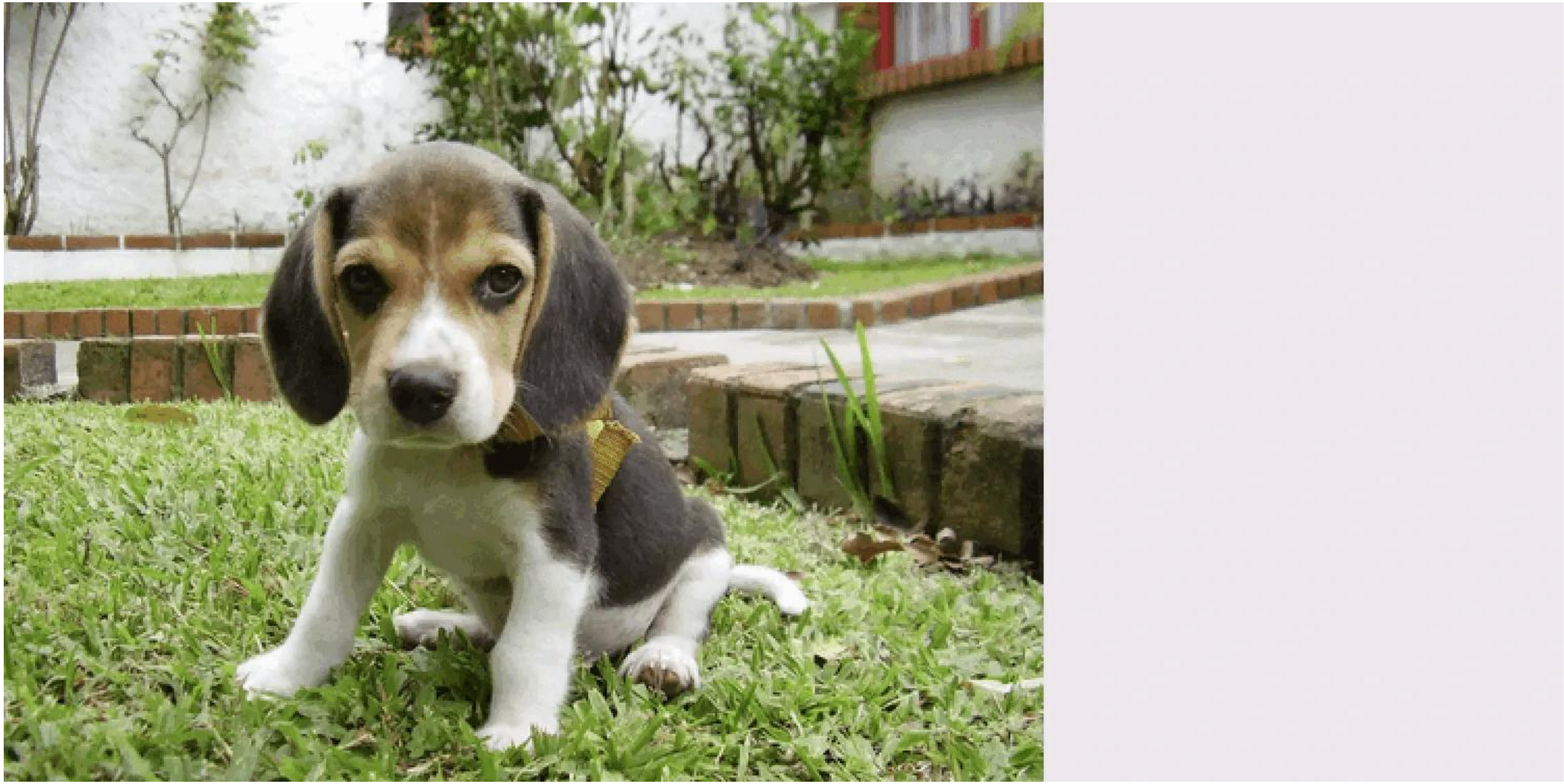


(b.1) Human Eva. w/o **BDG** or **ECG**



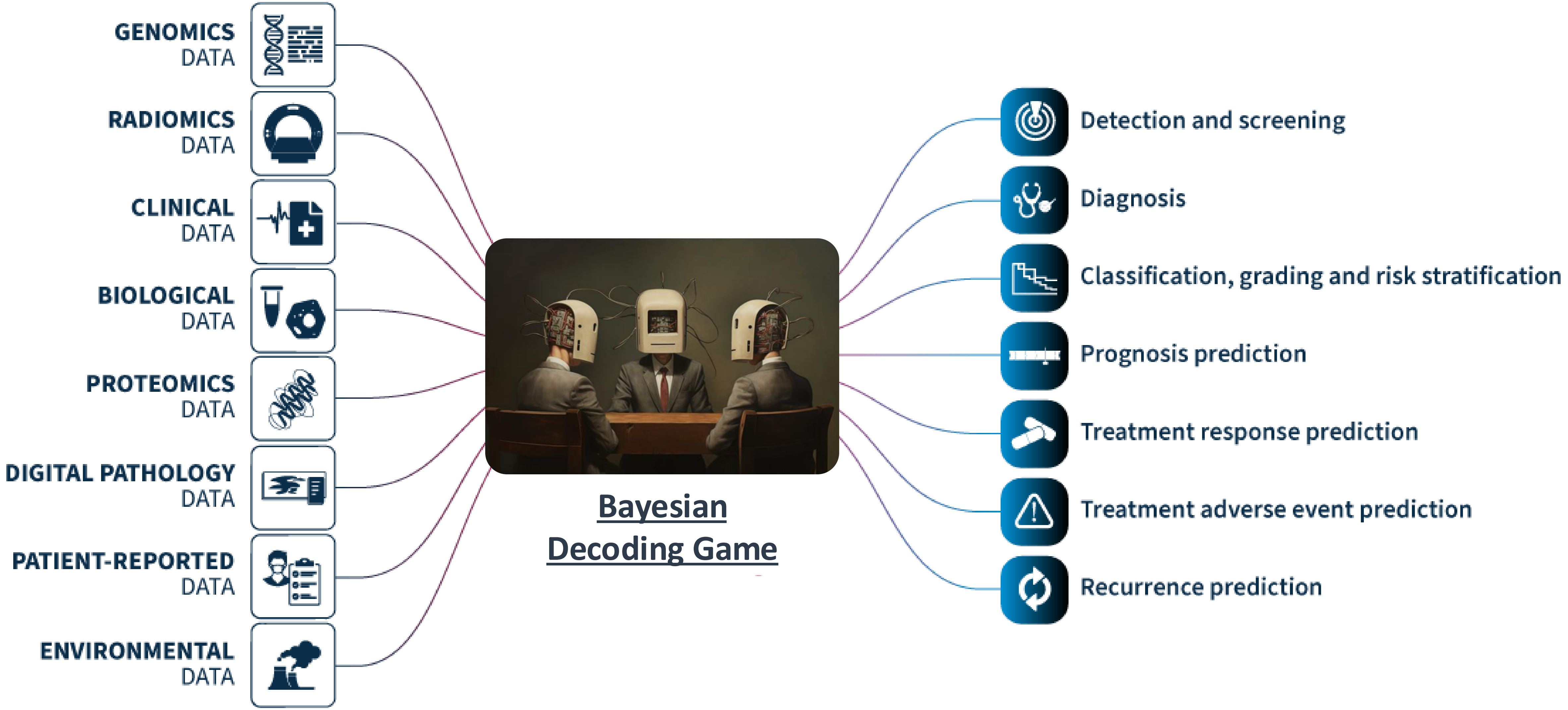
(b.2) Per-case report

~~Human vs. AI~~ Human (Symbolic) & AIs Consensus



[1] https://openai.com/index/debate/

~~Hallucinations~~ Multi-modality Consensus



[1] https://www.sophiagenetics.com/science-hub/the-power-of-multimodal-data-driven-medicine/
[2] https://cbmm.mit.edu/news-events/news/multi-ai-collaboration-helps-reasoning-and-factual-accuracy-large-language-models

Thank you for your attention.