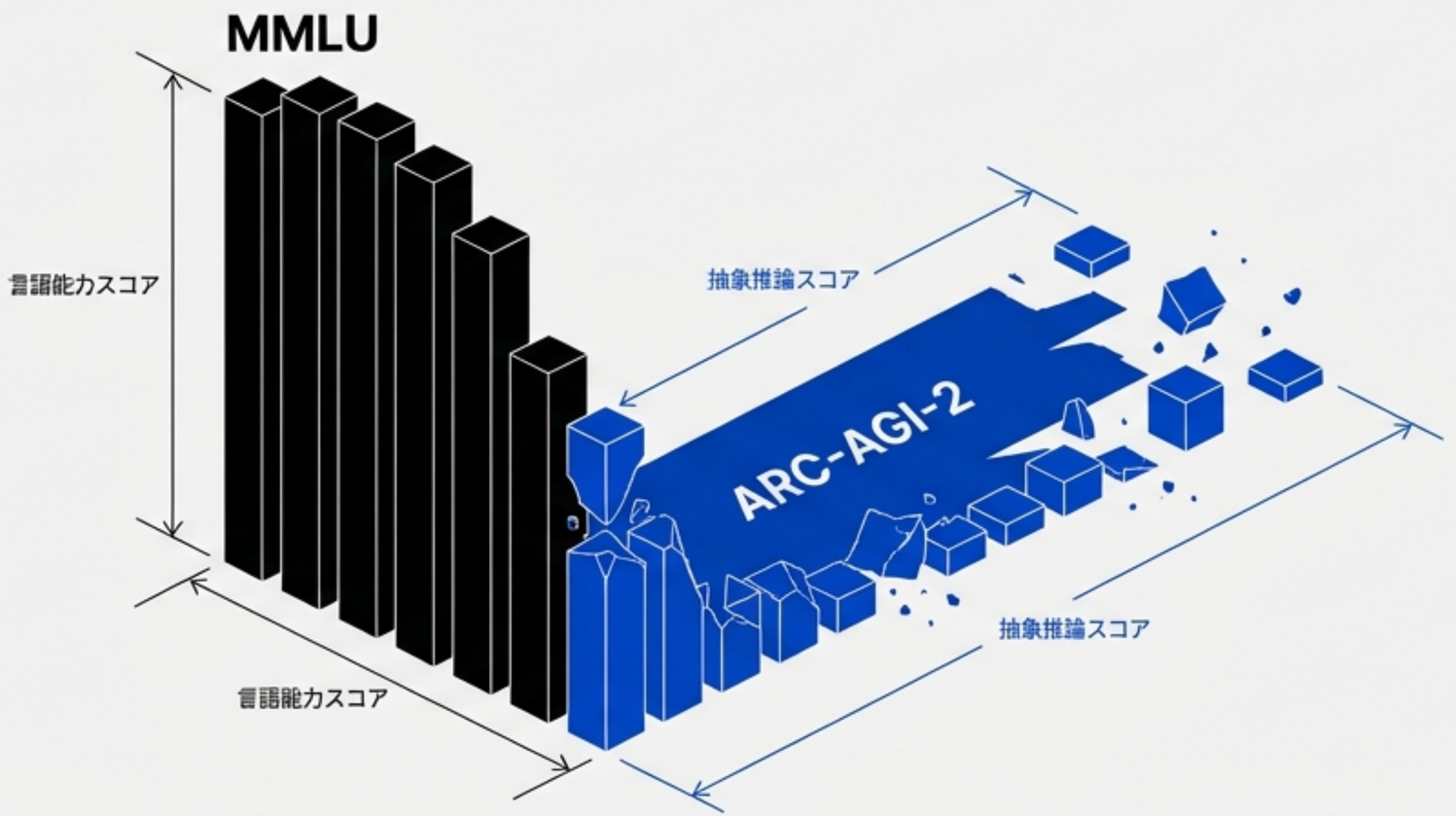
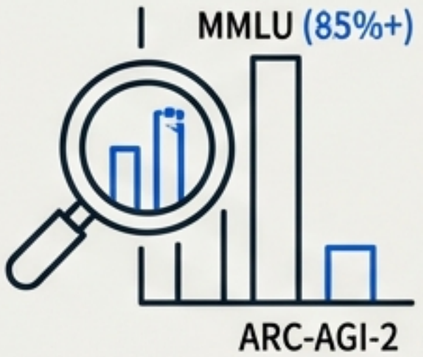
中国製LLMとARC-AGI-2：ベンチマーク乖離の構造分析

高い言語能力と低い抽象推論スコアの背景にある「評価の罠」と「技術的解法」



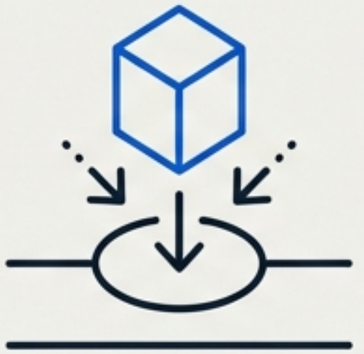
エグゼクティブサマリー：知能の欠如ではなく「最適化の不一致」

Observation (現象)



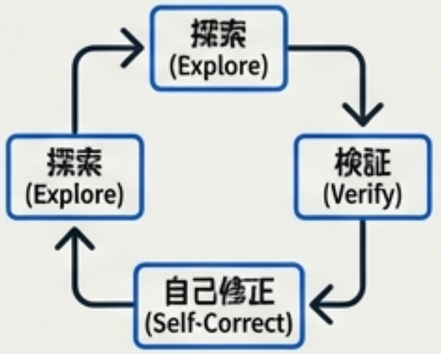
中国製LLM（Qwen, DeepSeek等）は知識タスク（**MMLU 85%+**）で世界最高峰だが、ARC-AGI-2では構造的制約によりスコアが低迷する傾向にある。

Root Cause (原因)



ARCは「確率的テキスト生成」ではなく「**ピクセル単位の厳密な一般化**」を要求する。知識への依存を極小化し、厳格な**pass@2** (2回のみの試行) 制限が、従来の試験対策 (暗記・パターン学習) を無効化している。

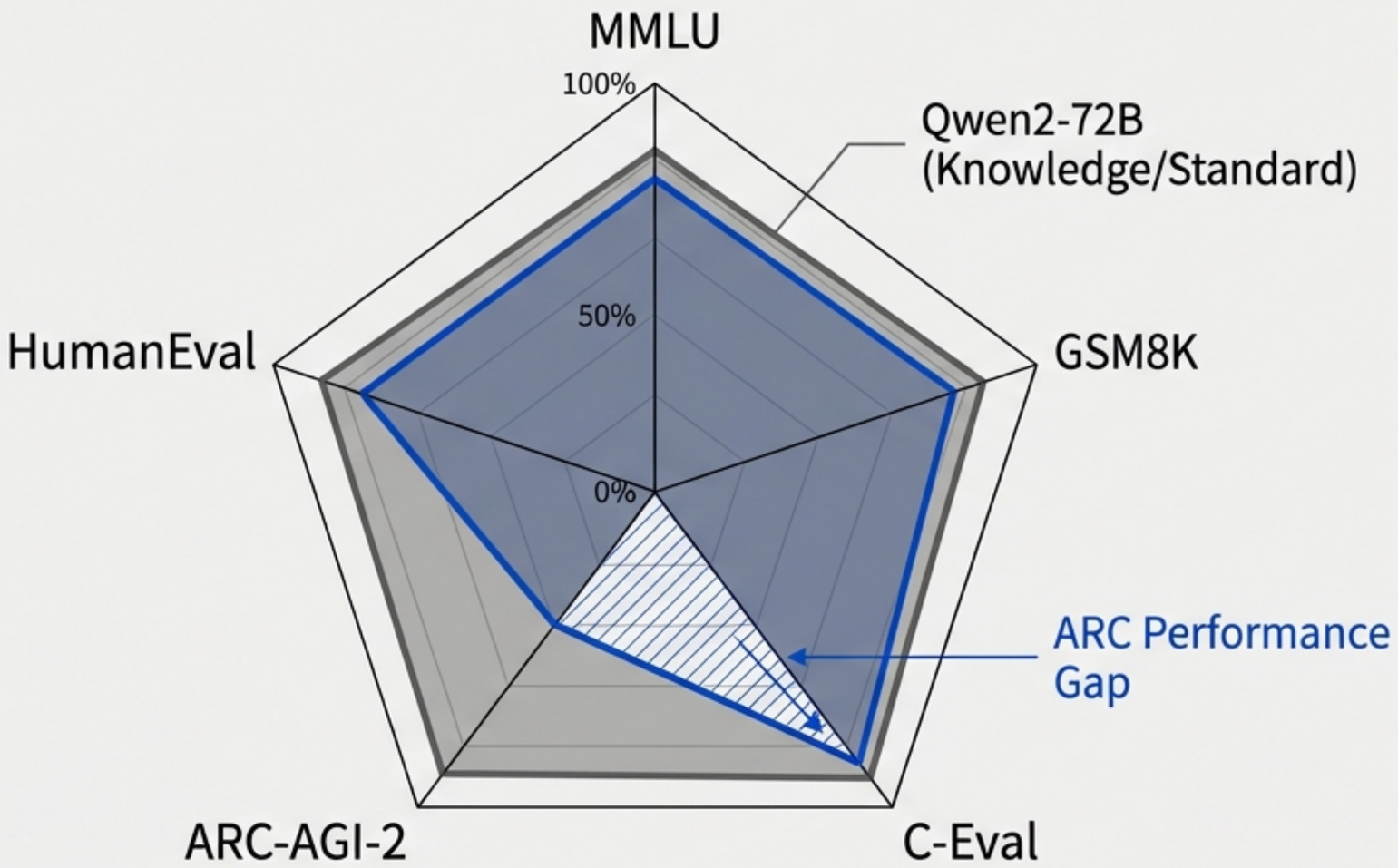
Solution (解決策)



解決策はモデルの大規模化ではなく、推論プロセスへの**「System 2」アプローチ**（探索・検証・自己修正ループ）の実装にある。

2026年初頭の検証済み最高スコアは、このアプローチにより77.1%に達している。

現象の定義：高スコアのパラドックス

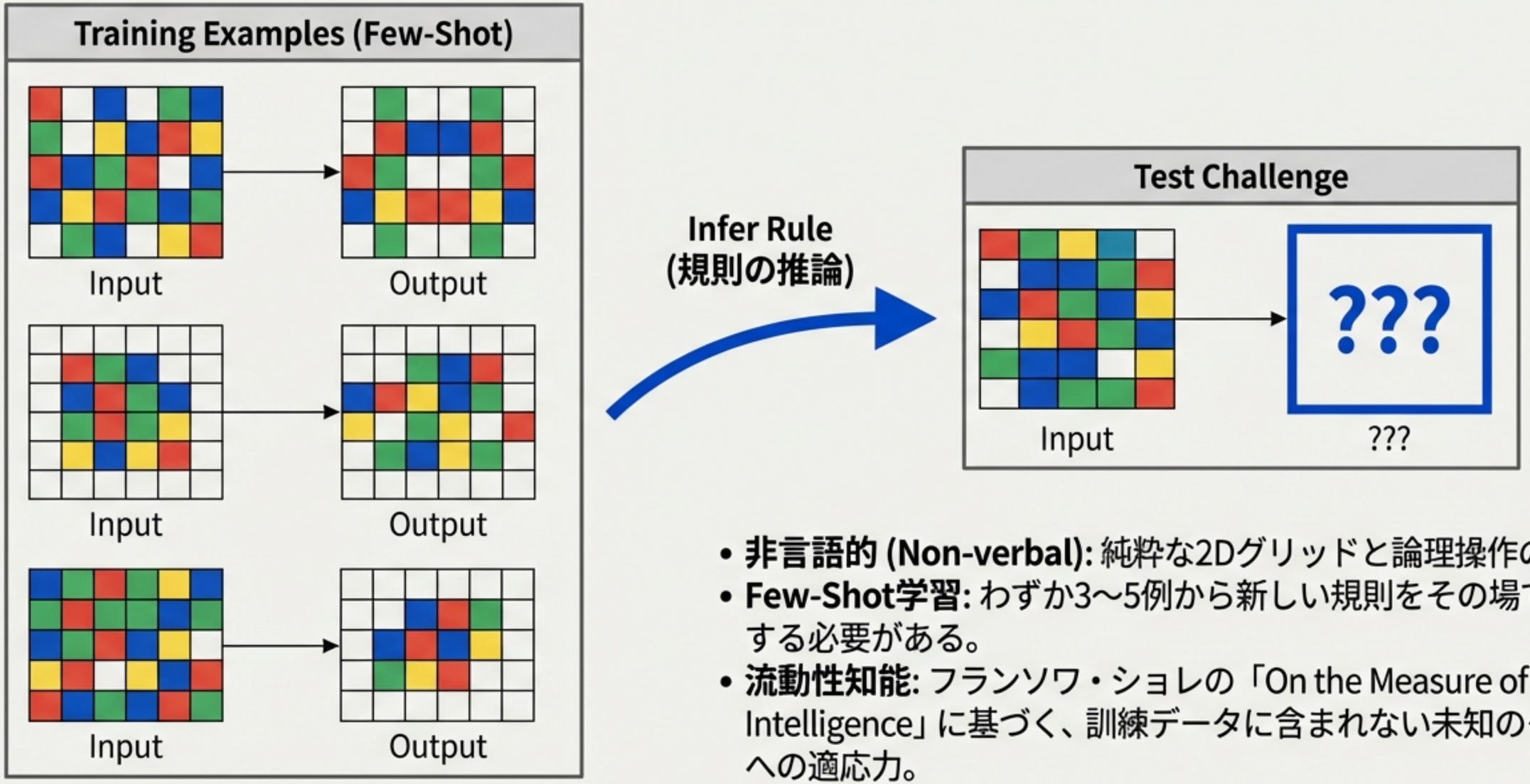


「世界クラスの知識」は「未知のパズルへの適応力」を保証しない。

Qwen2-72BやKimi K2等は、MMLUやHumanEvalでSOTAクラスの成績を示すが、ARC-AGI-2の「純粋なLLM単発推論」では0%〜数%に留まるケースが多い。

Source: Qwen2 Technical Report / ARC Prize Initial Analysis (March 2025)

敵を知る：ARC-AGI-2の正体



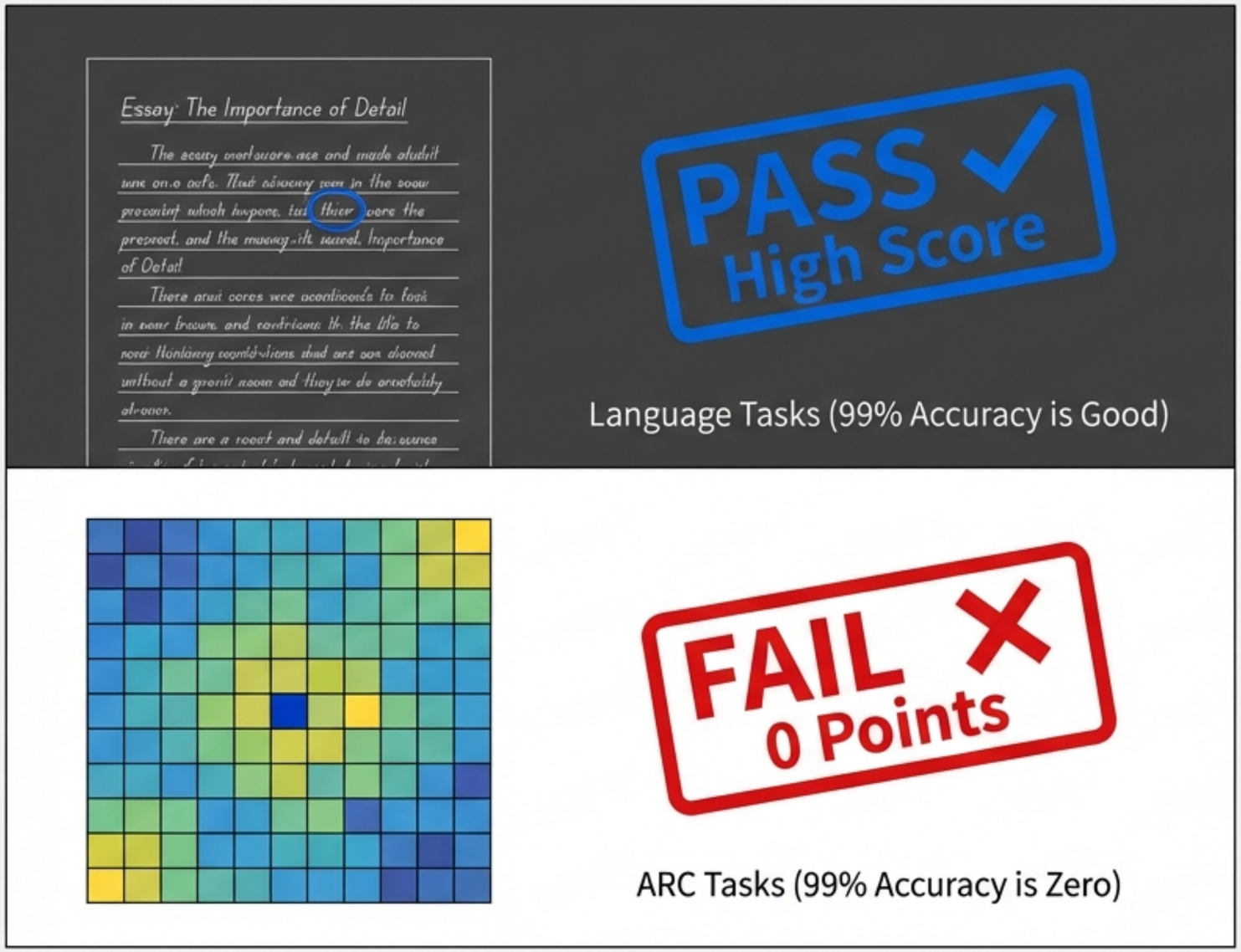
- **非言語的 (Non-verbal):** 純粋な2Dグリッドと論理操作のみ。
- **Few-Shot学習:** わずか3〜5例から新しい規則をその場で学習する必要がある。
- **流動性知能:** フランソワ・ショレの「On the Measure of Intelligence」に基づく、訓練データに含まれない未知のタスクへの適応力。

構造的ミスマッチ：「試験巧者」が失敗する理由

Feature	Standard Benchmarks (MMLU)	ARC-AGI-2
Metric（指標）	**確率的** (Probabilistic)	**決定論的（Deterministic）**
Success Criteria（合格基準）	pass@k (Multiple Attempts) / Partial Credit	**pass@2 (Strict Limit) / Exact Match (0 or 1)**
Content（内容）	Knowledge-heavy, Text-based	**Logic-heavy, Visual/Spatial**
Optimization（最適化）	**過去問学習**（Leakage）や**形式慣れが有効**	**厳格なPrivateセットにより、暗記が通用しない**

中国製LLMは「Standard Benchmarks」に過剰適合しているが、
その**最適化は「ARC-AGI-2」**ではペナルティとなる。

厳密一致の壁：部分点なき世界

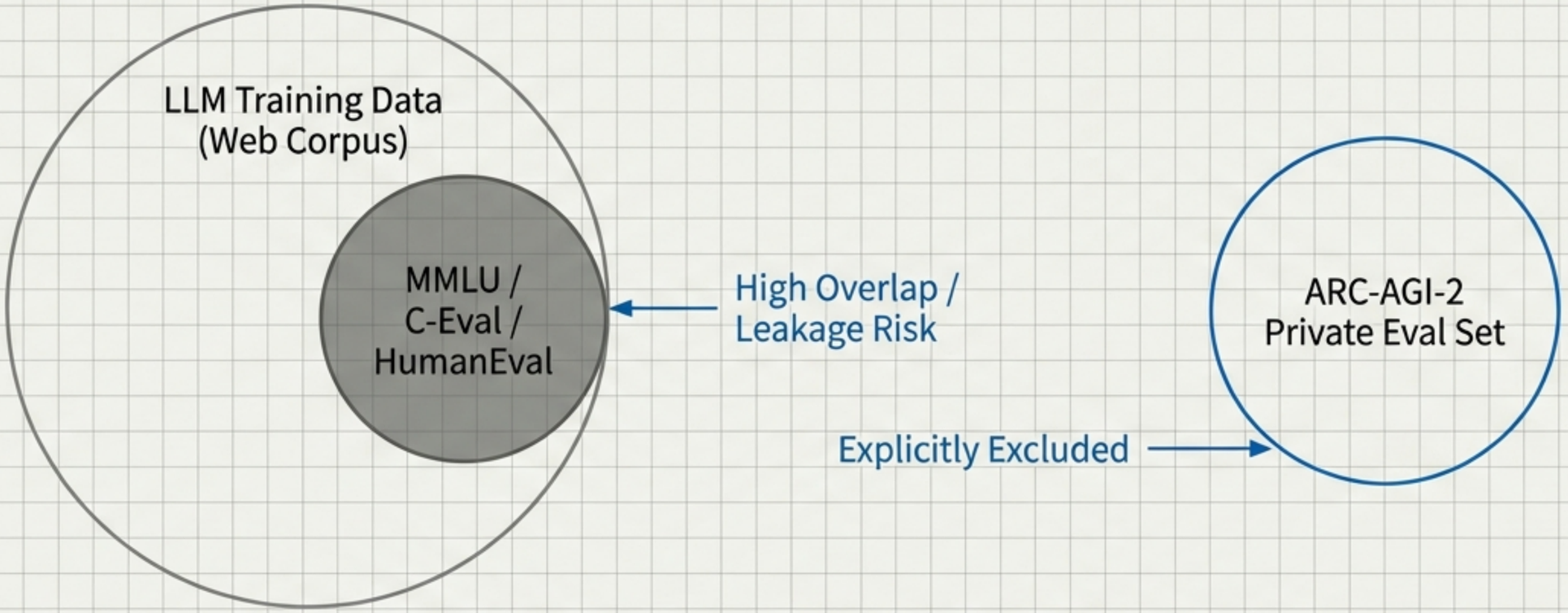


The Issue: 生成モデルの「だいたい合っている」能力は、ARCでは無価値となる。

Impact: 学習シグナルが極めて粗い。モデルは「惜しい」ことに気づけず、修正のフィードバックループが回らない。

Constraint: 競技ルールでは pass@2（２案提出）しか認められず、数撃ちゃ当たる（pass@100）戦略は封じられている。

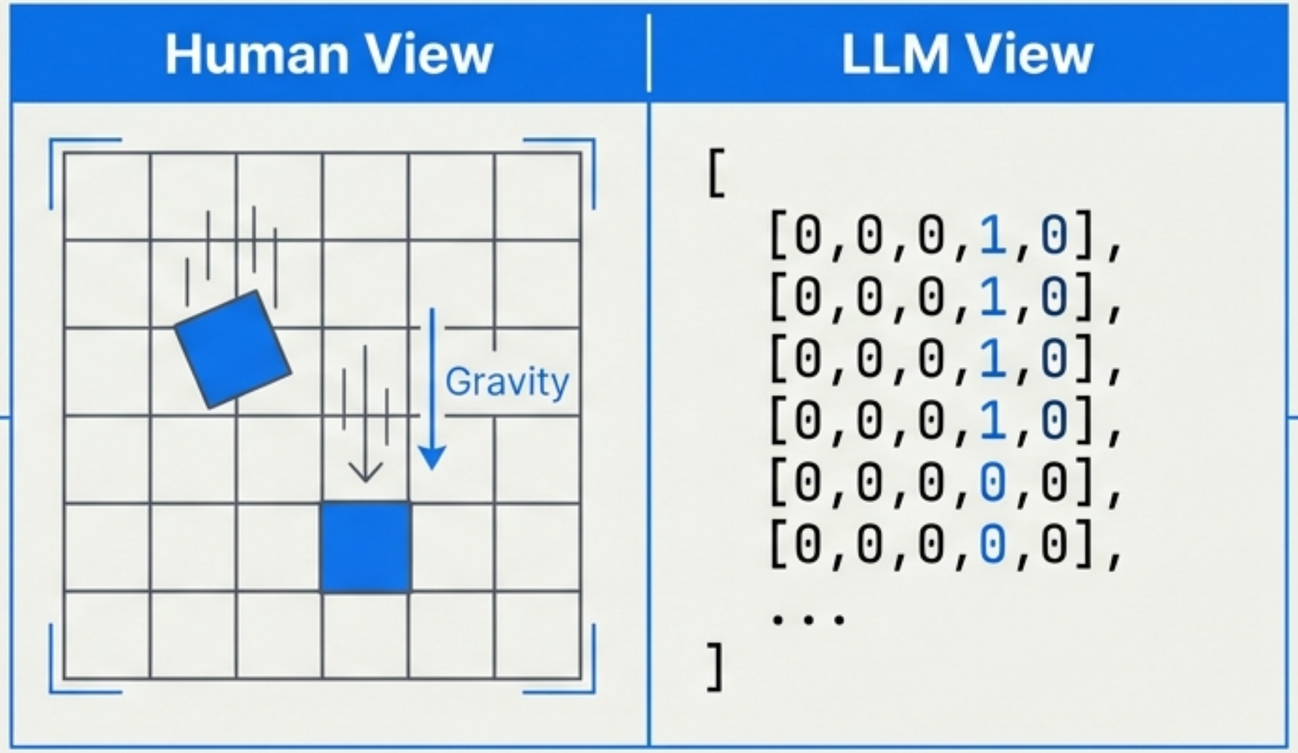
「試験対策」が通用しない：データセットの分離



- MMLU等の公開ベンチマークは、訓練データへの混入（汚染）や形式への過学習が起きやすい。
- ARC-AGI-2は「Semi-Private / Private」セットにより、未知のタスクのみを評価するよう設計されている。
- Insight: 高いMMLUスコアは「能力」よりも「データの記憶」を反映している可能性があり、その優位性はARCでは消失する。

仮説A：表現の壁 (Representation Gap)

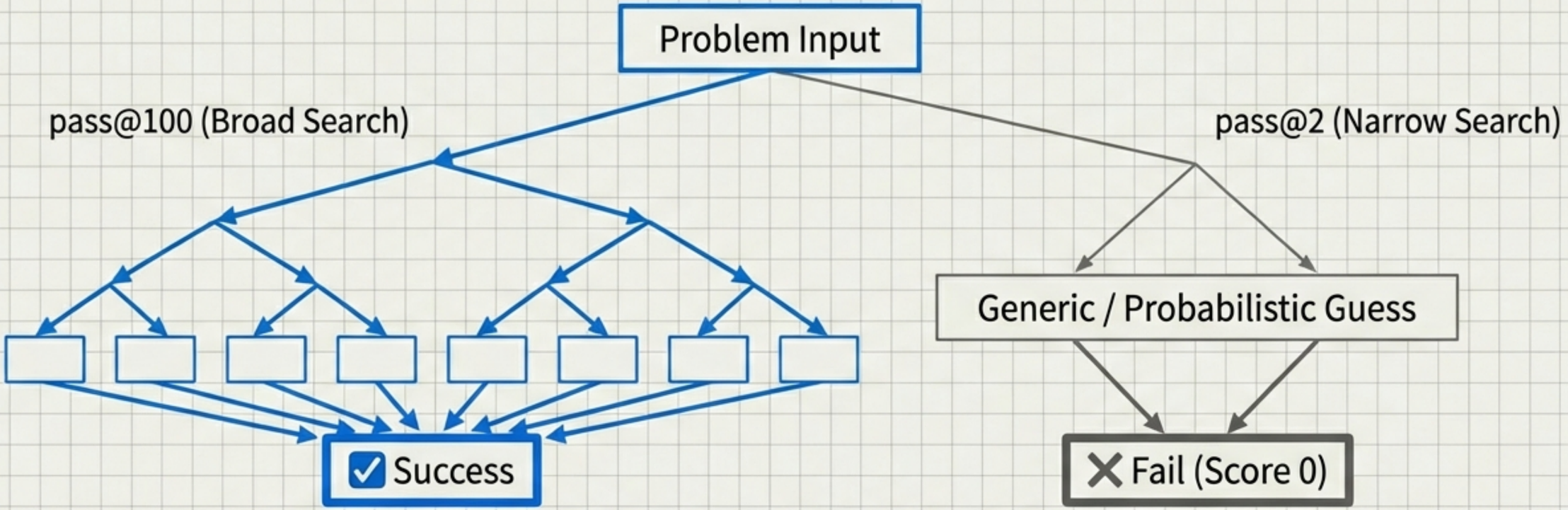
Spatial Inductive Bias
(Gravity, Geometry)



1D Token Stream

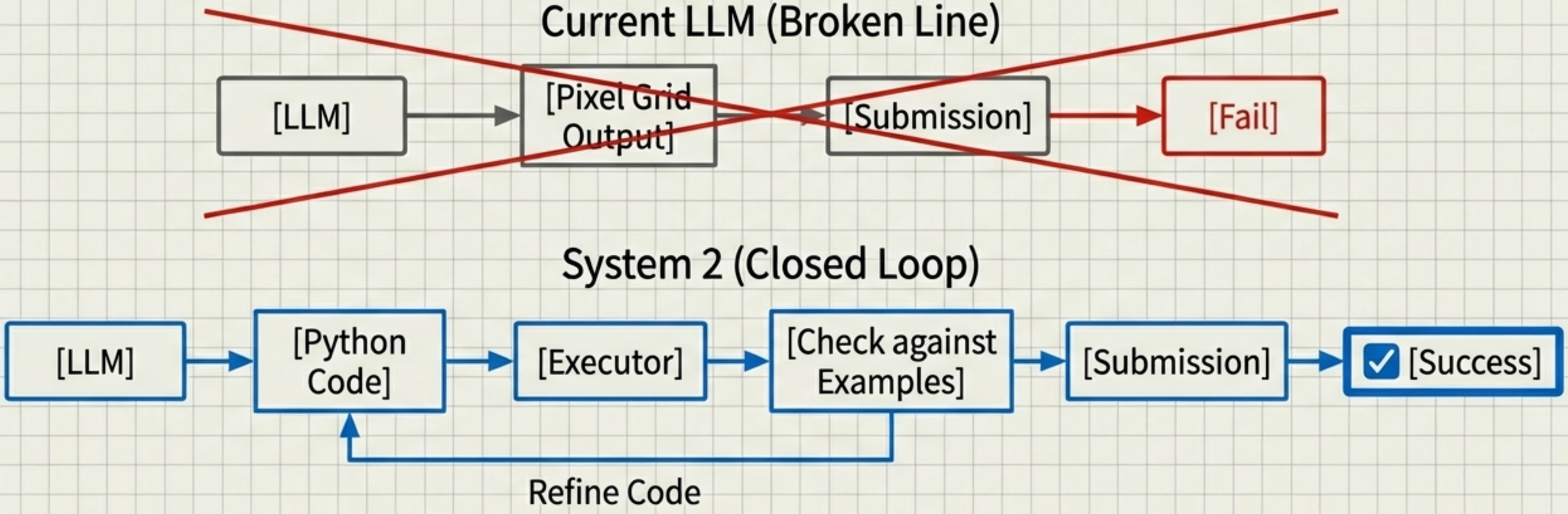
- テキストベースのLLMは、回転・対称性・重力といった「空間的帰納バイアス」を先天的に持たない。
- 2Dグリッドを1Dのトークン列として処理するため、視覚的には自明な規則（例：右に90度回転）の発見に膨大な計算リソースを浪費する。

仮説B：探索の不足 (Search Constraint)



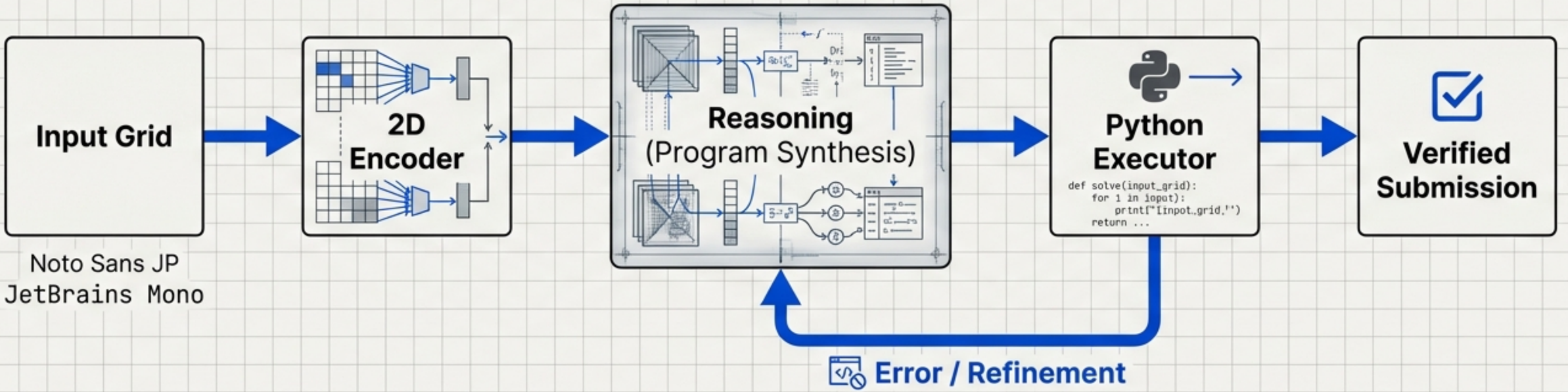
- LLMは確率的に最も「ありそうな」答えを出すが、ARCの正解は確率分布の「外側」にある特異な規則であることが多い。
- Requirement: 異なる仮説（色ベースの規則 vs 形状ベースの規則）を意図的に探索する多様性が必要。

仮説C：検証ループの欠如 (Verification Gap)



- ARC Prizeの分析: Direct Prompting < 5% vs. Verification Loop Systems >> 20%
- Diagnosis: 中国製LLMの低スコアは、モデル単体の性能不足ではなく、「検証器（Verifier）」を伴うシステム設計の不在に起因する可能性が高い。

解決策：システム化された推論 (System 2 Architecture)



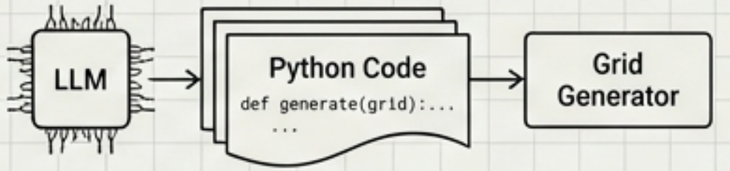
Noto Sans JP
JetBrains Mono

From Chatbot to Reasoning Engine: Don’t predict pixels, predict logic.

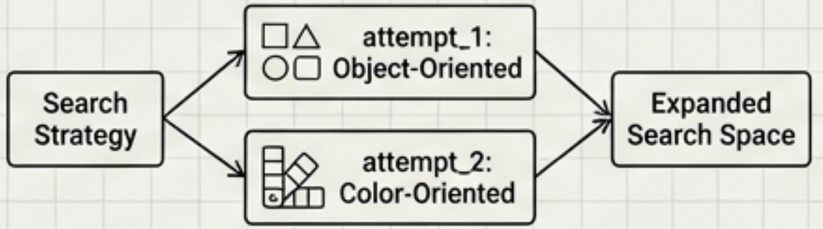
実装への提言：エンジニアリング・ロードマップ

ENGINEERING ACTION ITEMS / TECHNICAL MANIFEST

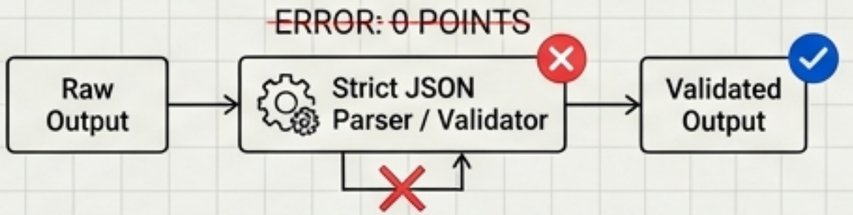
- [x] **Code over Pixels:** 直接グリッドを予測するのではなく、グリッドを生成するPythonコードを予測させる。



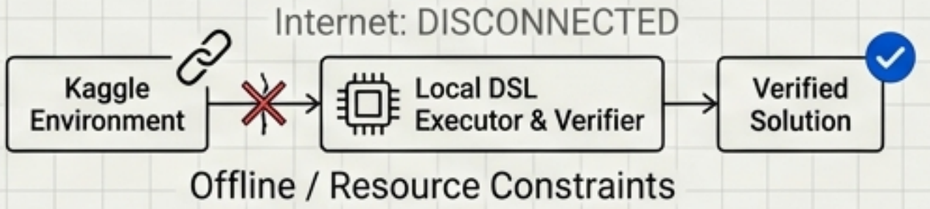
- [x] **Diversity in Search:** attempt_1とattempt_2で異なる戦略（例：物体指向 vs 色指向）を強制し、探索空間を広げる。



- [x] **Robust Formatting:** JSONパーサや厳格なフォーマット検証器を導入し、形式エラーによる「0点」を根絶する。



- [x] **Local Execution:** Kaggle制約（ネット切断・計算資源）を想定し、ローカル環境でのDSL実行・検証を実装する。

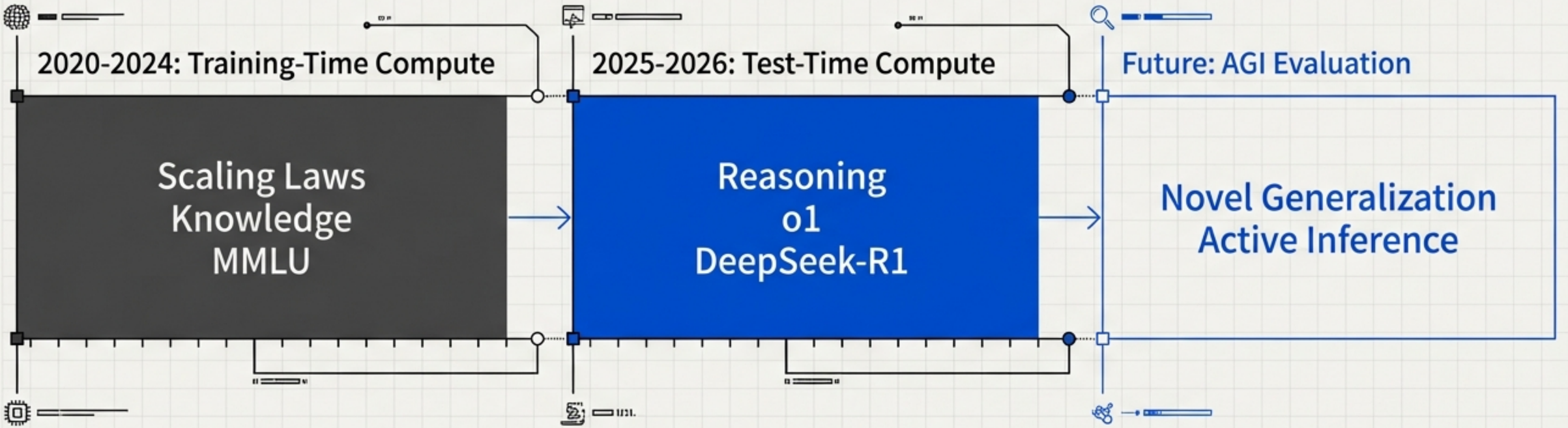


検証実験の設計 (Experimental Validation)

Variable (変数)	Options (条件)	Hypothesis (仮説)
Prompt Language	Chinese vs English	Minimal Impact (Task is non-verbal)
Representation	JSON vs Image/2D	High Impact on pass@2
Verification Loop	Direct Gen vs Code+Exec	**Highest Impact (Critical for exact match)**

Goal: 「国籍（中国製）」が原因ではなく、**「手法」が原因**であることを定量的に証明する。

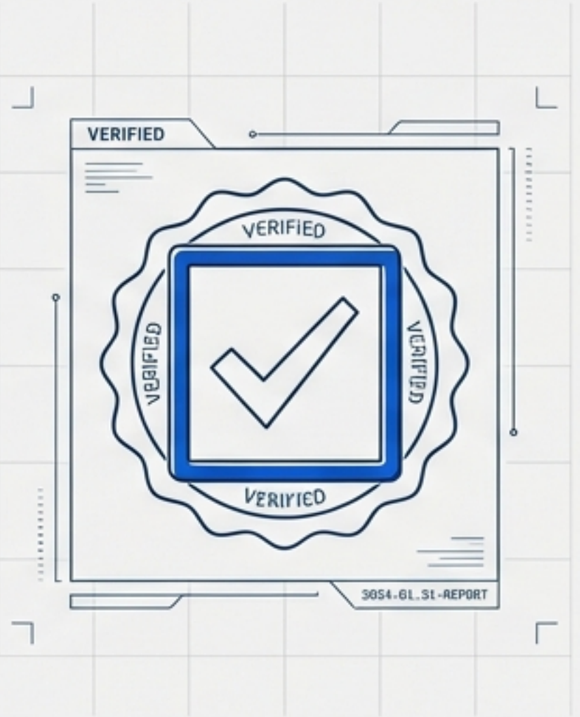
今後の展望：スケーリング則の先へ



ARC-AGI-2の攻略は、単なるベンチマーク向上ではなく、モデルを「記憶装置」から「思考装置」へ進化させるための重要なマイルストーンである。

結論と提言 (Conclusion & Recommendation)

1. • ベンチマーク乖離は「失敗」ではなく、評価指標の「進化」である。
2. • MMLU等の高スコアは「知識最適化」の結果であり、ARCが測る「未知への適応」とは異なる。
3. • 成功の鍵は、**探索・検証・適応（Search, Verify, Adapt）**を組み込んだシステム設計にある。



Action: Shift focus from pre-training optimization to inference-time reasoning architectures.