[AUDITED: 2026-05]

独立監査レポート：ChatGPT-5.2 Thinking 東大・京大首席合格主張の検証

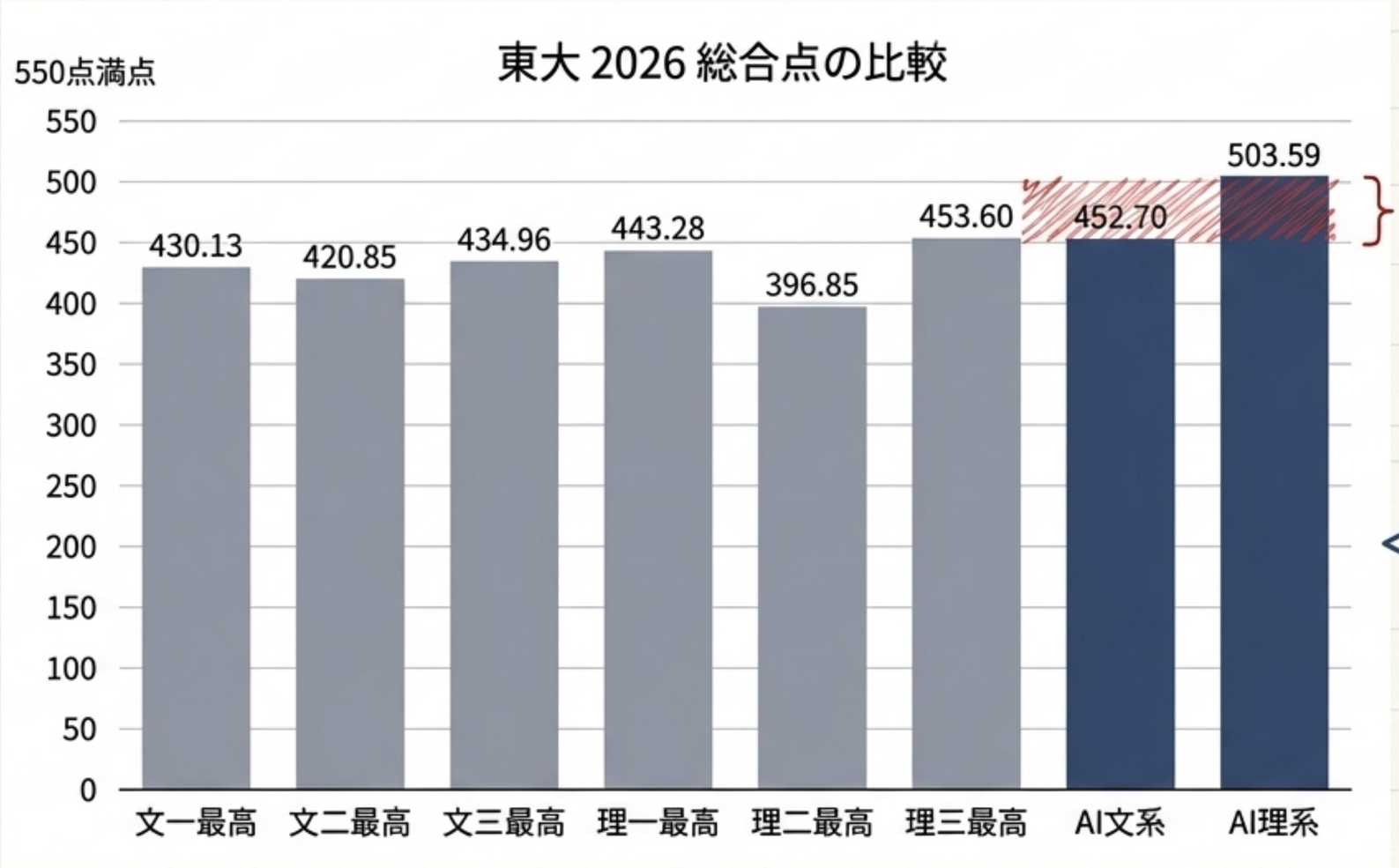
公開一次資料から読み解く、AIの実力とPRの境界線

MAY 2026 / THE AUDIT REPORT

The Verdict: 監査結論のサマリー

東京大学（理三首席相当・数学満点）	PRの主張: 50点差をつけての首席相当。 監査結果: 公開数値（公式最高点 453.60 vs AI得点 503.59）の突合で完全に整合。	[VERIFIED] 丸め誤差の範囲で事実
京都大学（全19学部・学科で首席超え）	PRの主張: 医学部を含む全学部で最高点を突破。 監査結果: 医学科は確認可能だが、他学部の独自換算尺度がブラックボックス。基準年度のブレも存在。	[PARTIALLY VERIFIED] 一般化には証拠不十分
手法の再現性（学術的証明）	PRの主張: 人間と同条件での実力証明。 監査結果: プロンプト全文、モデル設定、採点票が非公開。別日程のスコアの合成成績。	[UNVERIFIED] 第三者による厳密再現は不能

The Evidence: 東京大学における「圧倒的スコア」の裏付け



大学公式の2026年度最高点と、公開された550点満点総計を同尺度で照合。

- 理科三類公式最高点: 453.60点
- AI 理系総計: 503.59点

結論: 報道された「50点上回る」という主張は誇張ではなく妥当な丸め処理。公開値ベースにおいて、東大全6科類での首席超えは強く裏付けられる。

The Discrepancy: 24時間で動いた京都大学の「基準点」

Day 1	Day 2
2026-04-27（note公開）	**2026-04-28（プレスリリース・報道）**
基準点: 2025年 医学科最高点	**基準点: 2026年 医学科最高点**
1105.87点	1098.25点
AI 1176.25点	AI 1176.38点
差分: +70.38点	**差分: +78.13点**

基準点がスライド（下落）

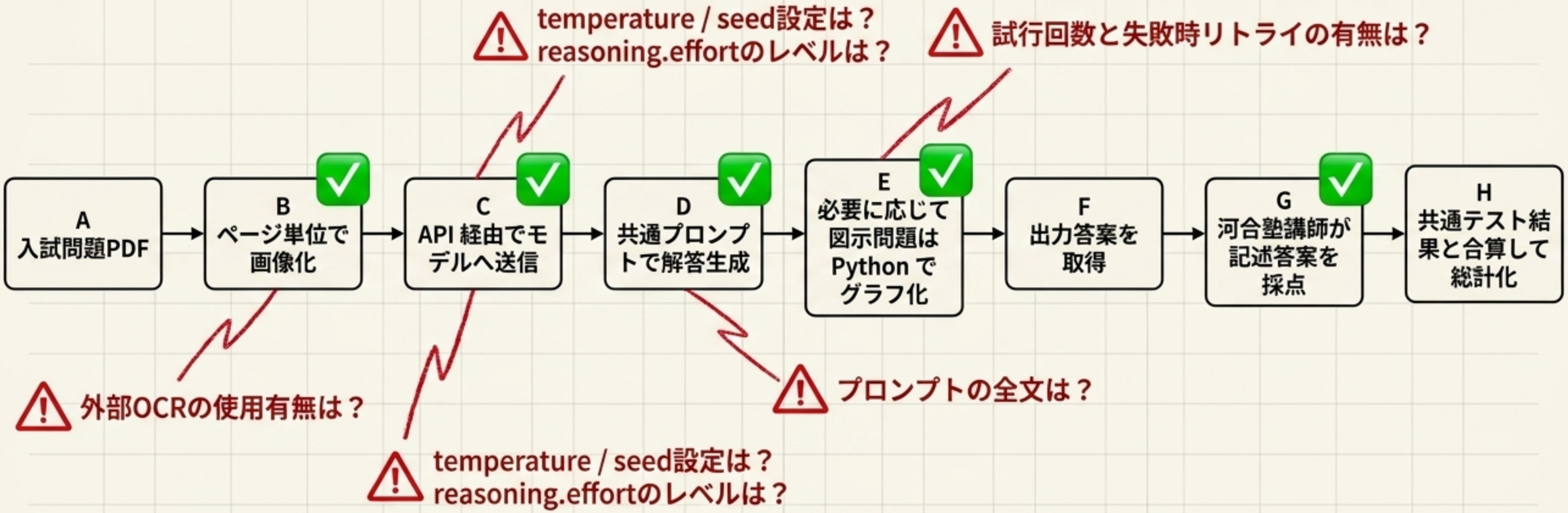
監査上の指摘: 医学科の高得点自体は確認できるが、主張の根拠となる数値が固定的ではなく、わずか24時間でより強い主張（リード幅の拡大）へ微修正されている点は監査上の減点要素である。

The Scope Gap: 「全19学部首席超え」はなぜ未検証なのか

- プレスリリース内で「ほぼ全学部」と「全19学部」で射程が揺れている。
- 検証可能（医学科のみ）：共通尺度で比較が可能。
- 検証不能（他18学部・学科）：AI側の換算値は存在するが、大学公式の総点表と直接突合できる「換算式」が説明されていない。

結論: 「全19学部首席超え」は現段階では第三者による再現・監査が不可能な「未検証の主張」である。

Anatomy of a Black Box:
第三者検証を阻む「隠された変数」



結論: 公表されたフローは明確だが、厳密な再現に必要なパラメータが欠落している。
これらが非公開である以上、学術的・計量的な意味での「確証」には至らない。

Changing Variables: 「1年で劇的な点数上昇」の真実

[2025年検証: ChatGPT o1]

東大理系総計: 374点

- ・プロンプトなし
- ・スクリーンショットを そのまま入力
- ・手動での文字起こし（国語）

純粋なモデルの進化？
それとも実験設計の最適化？

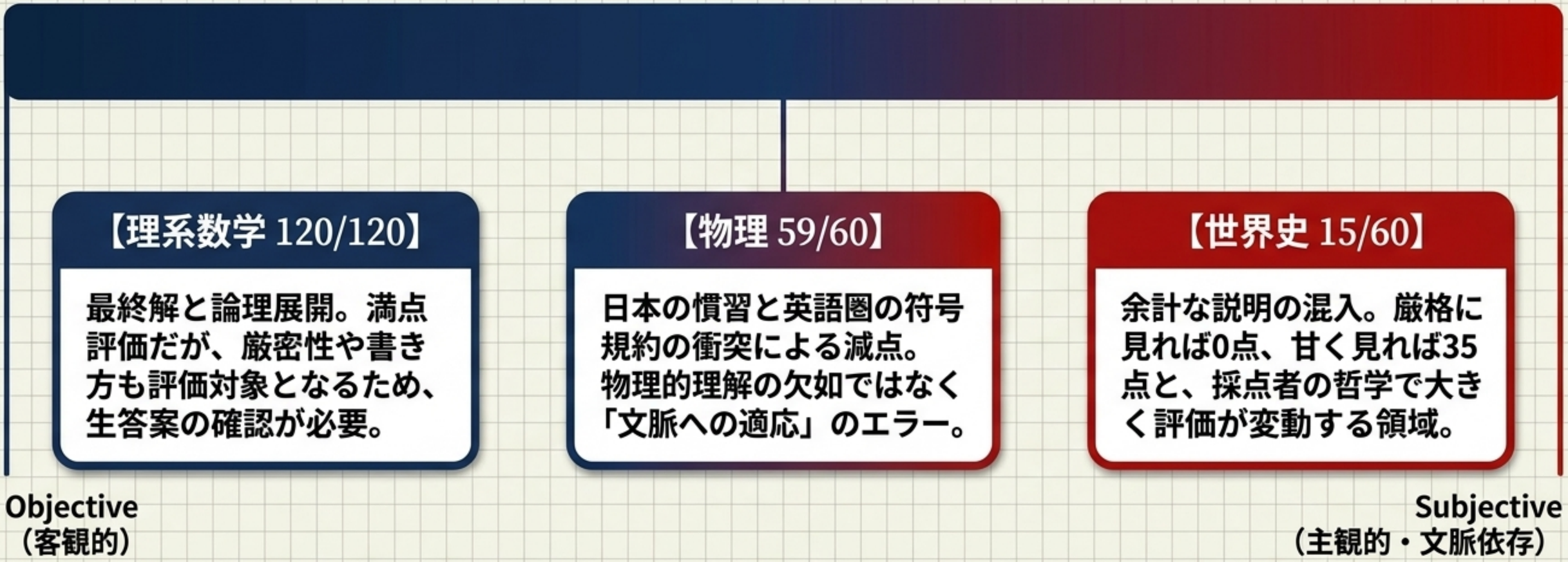
[2026年検証: ChatGPT-5.2 Thinking]

東大理系総計: 503.59点

- ・最適化された共通プロンプトあり
- ・自動受験システム構築
- ・図示問題でのPython連携

結論: 1年間で+129.59点の飛躍は事実だが、2026年検証には実装の最適化が多分に上乗せされている。点数上昇を「AIの進化」のみに帰属させるのは計量的に危険である。

The Subjectivity Spectrum: 「採点の揺らぎ」が生み出す誤差帯



監査上の見解:
河合塾講師による採点は合理的だが、inter-rater reliability（採点者間一致率）や二重採点のブラインド化プロセスが提示されていない。
公開された点数は「有力な専門家判断の推定値」であり、誤差帯の見積もりが欠如している。

The True Insight: AIは何をハックし、何につまずいたか

[Verified Capabilities: AIが圧倒した領域]

- 数理推論: 数学満点・理科高得点という圧倒的パフォーマンス

✓ レイアウト依存の読解:
図表・構造式の安定した認識力

✓ 外部ツールの活用:
Python連携による図示問題の突破

[Observed Limitations: AIが取りこぼした領域]

× 高度な日本語の含意:
比喩や皮肉の処理における脆弱性

× 論述構成力:
世界史の論理接続における大崩れ

× 文脈・慣習への従属: 物理の符号規約や、冗長な回答の制御不能

インサイト: 万能の知性が誕生したわけではない。試験が測る能力のうち、「複雑な定型知識と数理処理」に特化した強固なフロンティア・モデルの姿が浮かび上がる。

Reality Check: これは「入試の敗北」を意味するのか？

The AI Experiment (今回の検証)	Real Exam (人間の受験生)
・API経由の自動化システム	・紙の問題冊子とペンのみ
・共通テストと二次試験の別実験スコアを合算（合成成績）	・連続した過密日程による受験
・精神的・肉体的疲労ゼロ	・極度の精神的・肉体的疲労
・図示問題でのPythonコード実行	・厳格な時間制約と自力でのグラフ描画

結論：今回の総計スコアは、別日程で行われた実験の高得点を組み合わせた「合成入試成績」である。現時点で脅かされているのは「会場での筆記試験」そのものではなく、試験が測る能力の一部が「外部化・自動化可能」であることを証明したに過ぎない。

Recommendations: 今後の検証と教育への提言

For Universities (大学の適応)

- 採点基準（模範答案・減点例）の透明化。
- 一発の静的答案から、プロセスを問う「口頭試問・面接型」の導入。
- 高速計算より、未知の資料の咀嚼・批判を重視する設問設計へ。

For AI & Media (監査可能性の担保)

「首席級」を謳うための4要件の同時公開を徹底：

1. 生答案全文
2. 採点済み答案 / 採点票
3. 大学公式と突合可能な換算式
4. モデル設定・シード値

For Society (冷静な議論)

- 「AIが入試を解けたか」という二元論からの脱却。
- 「どの能力が自動化され、どの能力が人間中心に残るか」の緻密な分解と議論へ。

センセーショナルな見出しを越えて、事実に基づく「次なる教育と評価のアップデート」へ。