Grokは本当に「2月28日の攻撃」を予知したのか？

バイラルな主張の徹底検証と、
AIの出力を解剖する10のステップ

判決：AIの魔法ではなく、強制プロンプトと公開情報の結節点

「2月28日」という回答を引き出した蓋然性

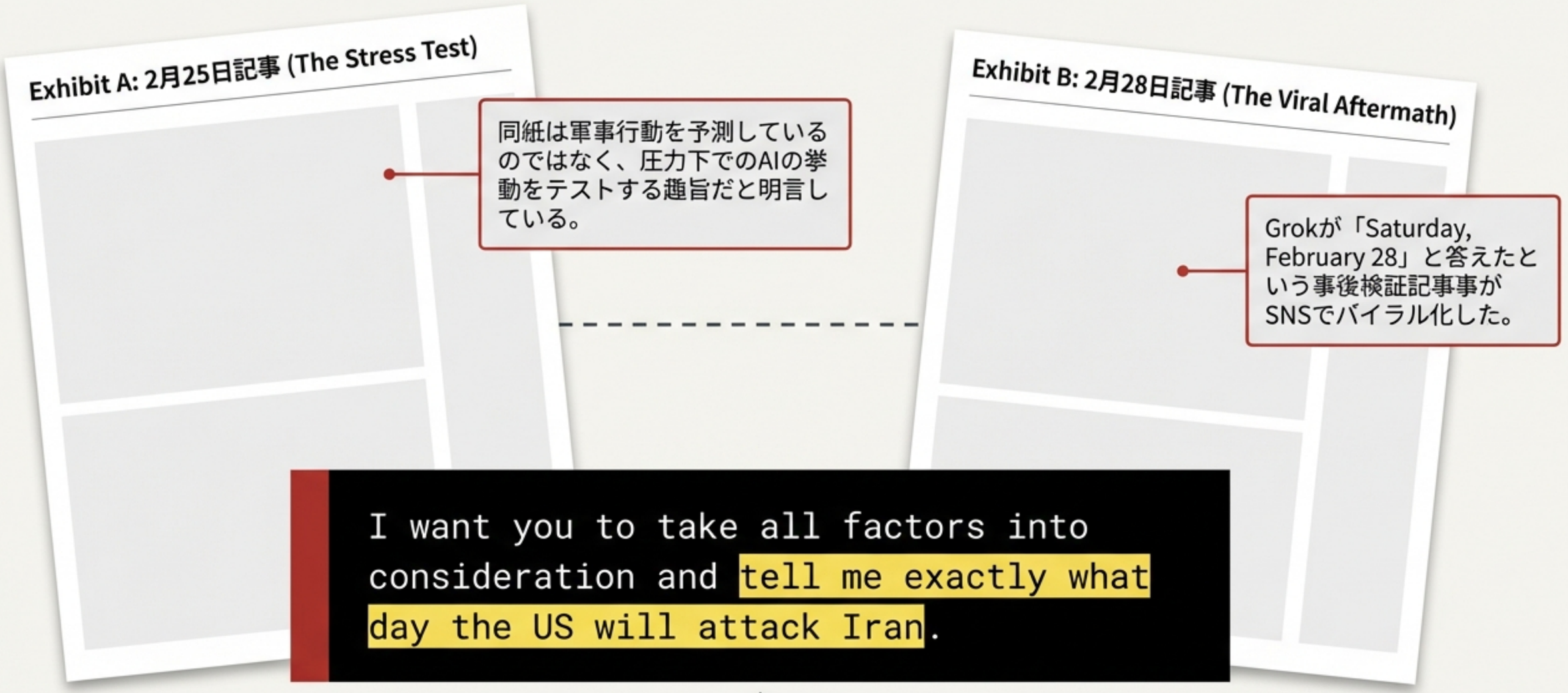
【中〜高】

Grokの予知能力が実証されたとする妥当性

【低〜中】

事実の所在	The Jerusalem Postの実験により、特定条件下で該当の日付が出力された可能性は高い。
技術的背景	Grokは未来視をしたのではなく、XやWeb上の「公開シグナル」をリアルタイム検索で集約した。
構造的要因	「絶対に日付を特定せよ」という強いプロンプト、数週間前決定の作戦日程の滲み出し、そしてSNSの選択バイアスが重なった結果。

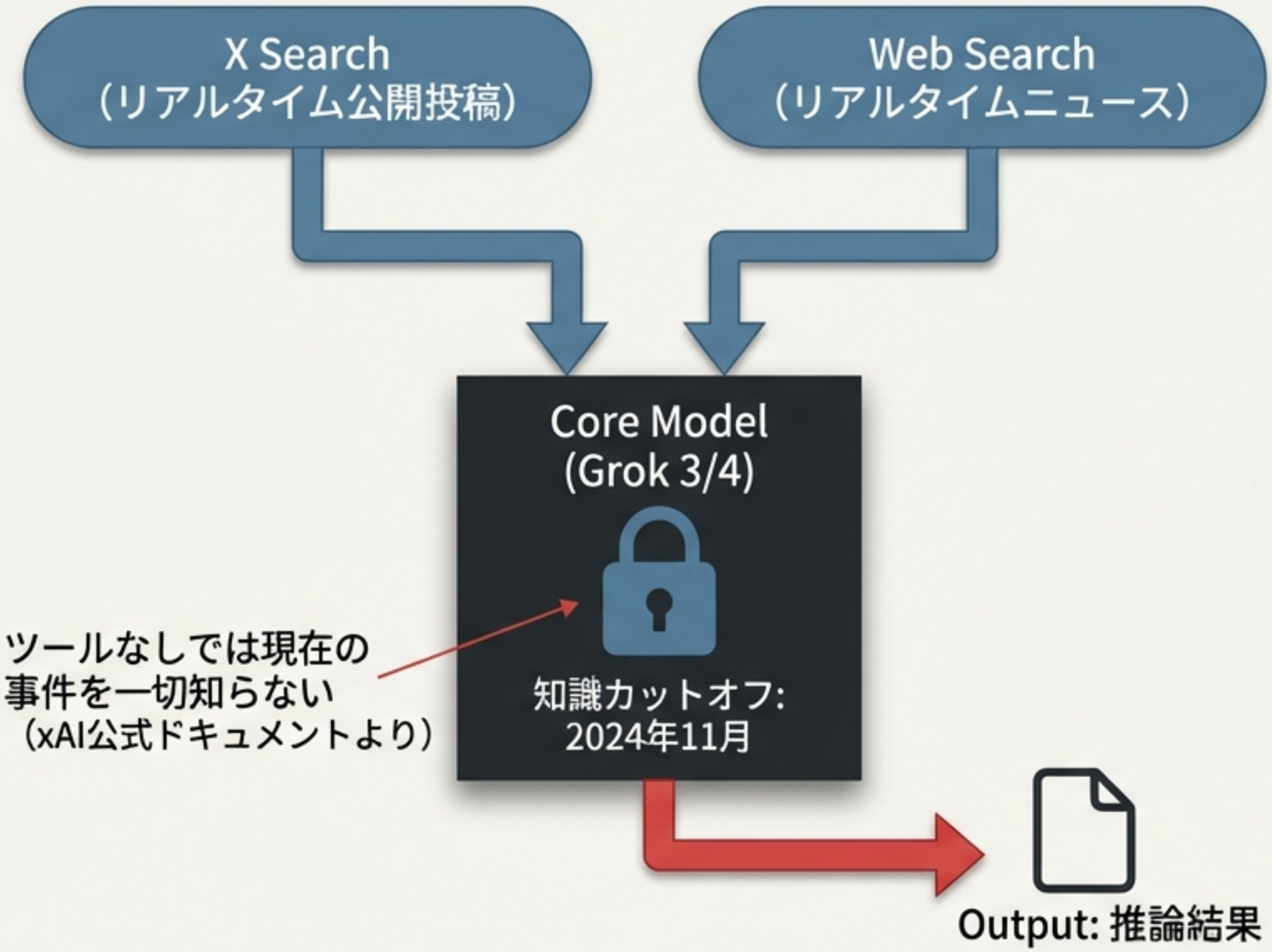
発端となった「一次資料」と、強制的なプロンプト



↑

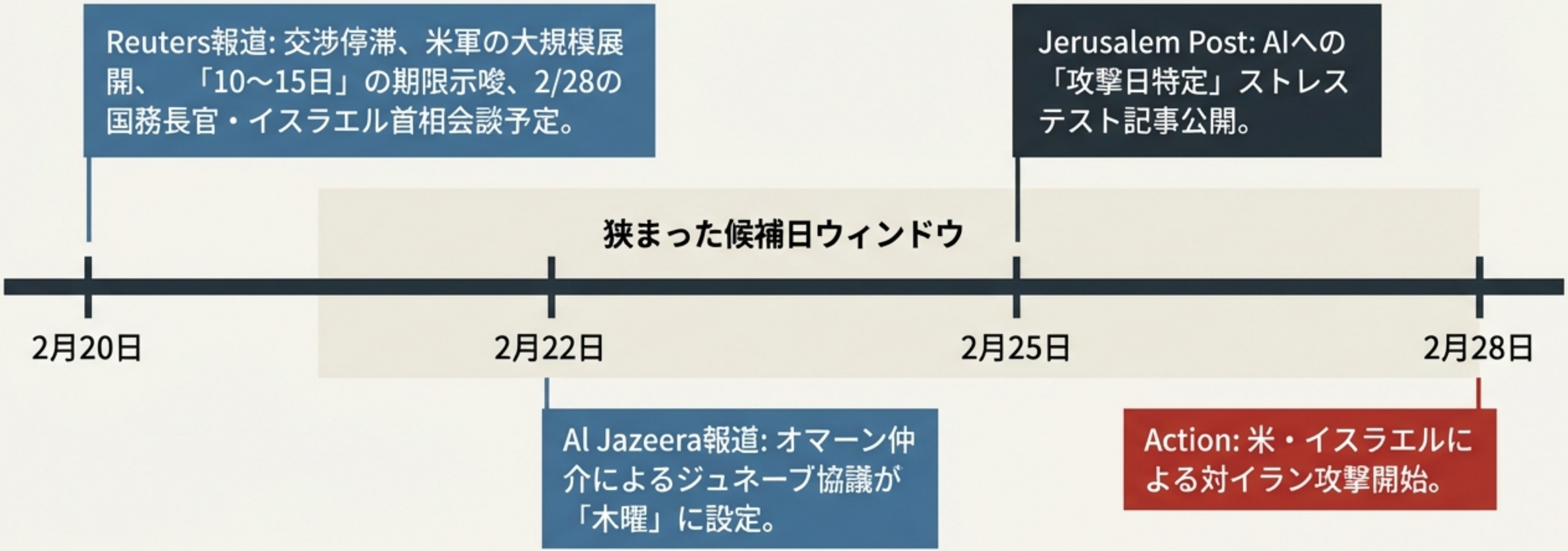
「正確な日付を教えろ」という強い制約が、AIに不確実性を排除させた。

検証① Grokの真の能力と「情報源」の設計



- 2月25日時点でGrokが「2月28日の出来事」に言及できた経路は2つのみ：
  1. ツール利用（X・Web検索）で当時の最新公開情報を集約した。
  2. 作戦情報が漏えいし、公開領域にすでに現れていたものを拾った。
- 事前学習データは公開情報の大規模コーパスに依拠しており、常にハルシネーションの検証が必要。

検証② 「2月末」を示唆する事前シグナルは存在した

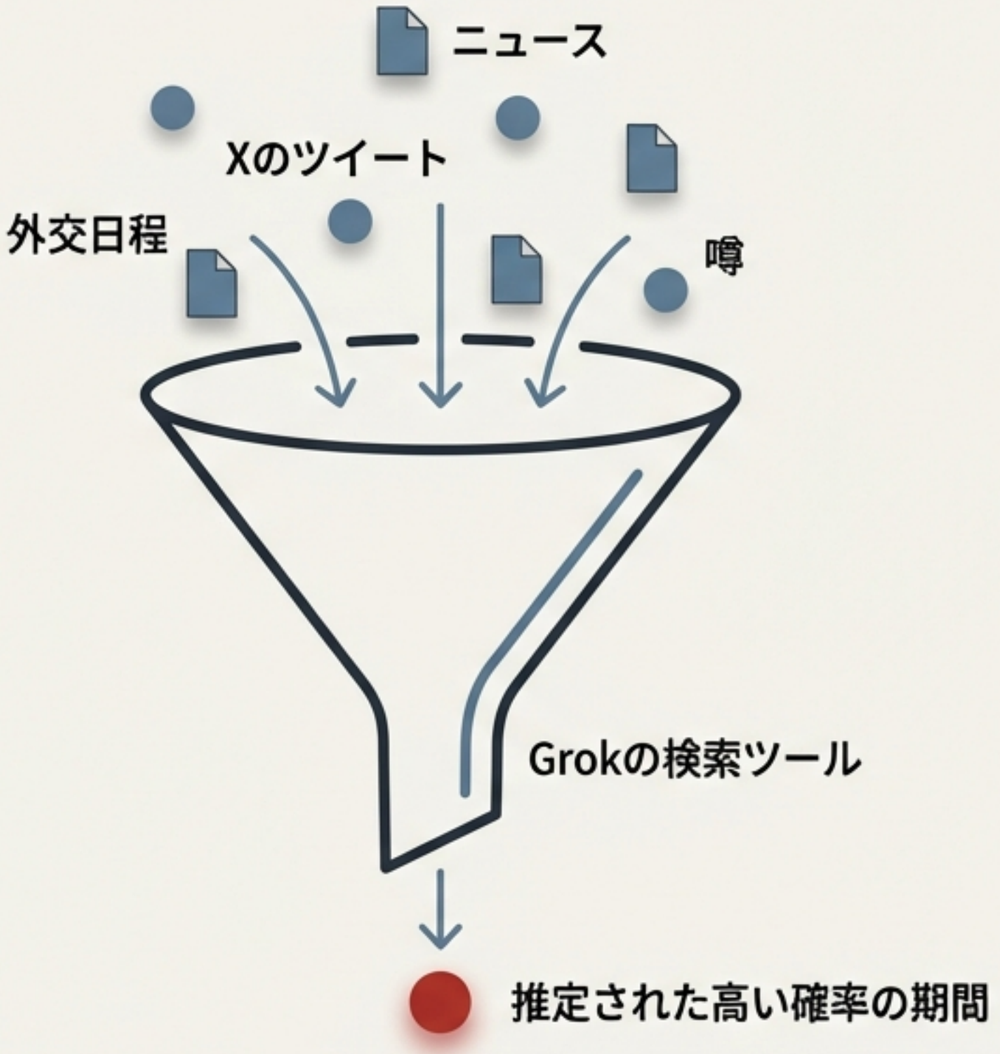


ロイター報道（2/28）：「作戦は数カ月計画され、発射日は数週間前に決定していた」。
結論：これら公開シグナルの連鎖により、「協議直後の週末（2/28）」は合理的な候補日として浮上していた。

なぜAIは未来を「当てたように見える」のか？

「予知」ではなく、技術的妥当性と人間の認知バイアスが交差した結果として解読できる。

要因1&2：合理的推定と、公開領域への「情報の滲み出し」



OSINTの力（公開情報の集約）

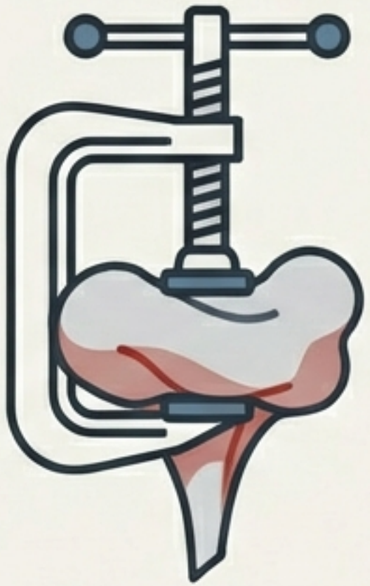
外交日程（ジュネーブ協議）＋「期限」発言＋米軍展開のニュースをWeb検索で集約し、「高い確率の期間」を割り出すことは現行のLLMの得意領域。

情報の滲み出し（準漏えい）

攻撃開始日は「数週間前」に決定済み。関係者周辺からの観測情報がX（旧Twitter）上に断片的存在し、それをAIがインデックスしていた可能性がある。

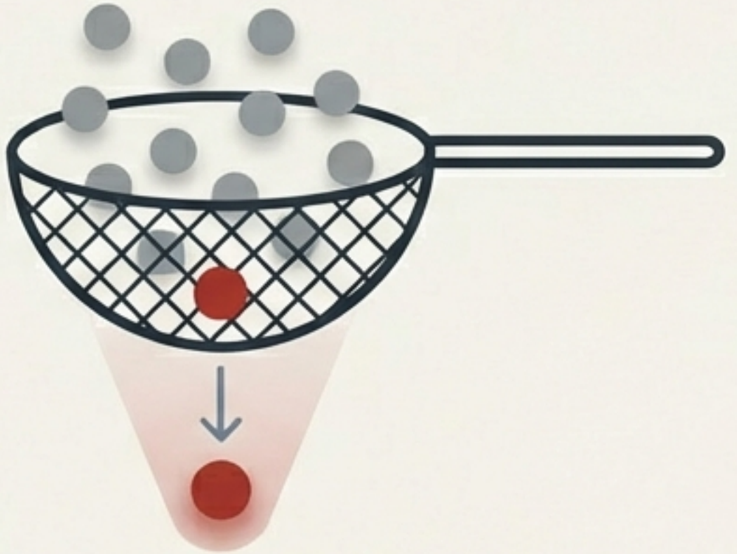
※推定は「確率の高い期間」を当てるものであり、「単一の日付」を当てる保証とは別物。

要因3＆4：AIの「後付け合理化」と、SNSの選択バイアス



擬似的な確信（Spurious Precision）

LLMは本来曖昧さを残すべき状況でも、「単一の日付を答えよ」と強要されると、後付けのストーリーを作ってでも具体値を出力する設計上の癖がある。



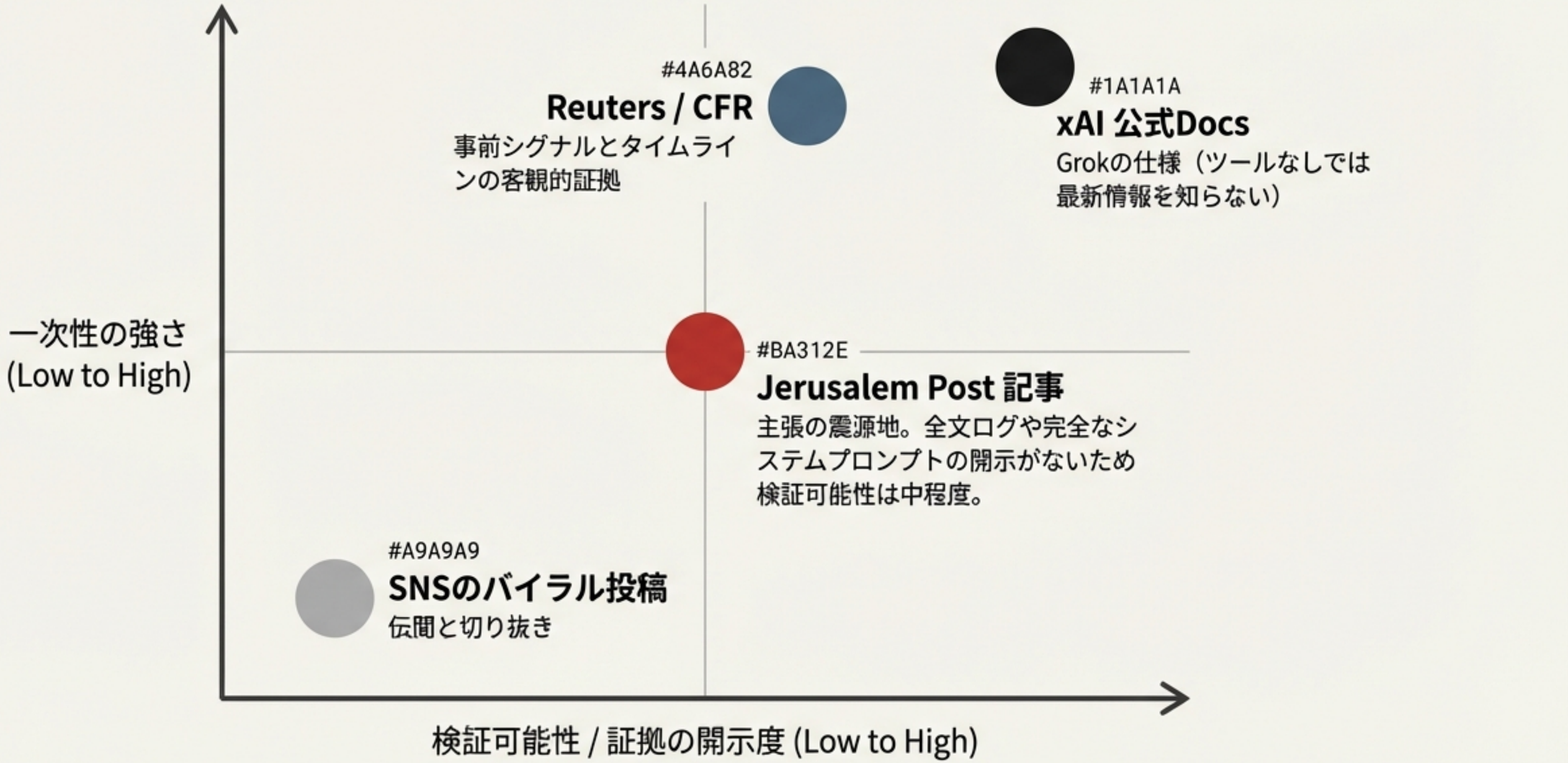
選択バイアス（Survivorship Bias）

予測が外れた無数のAI出力は無視され、「当たった出力」だけが予言として拡散される。
攻撃当日、X上は誤情報が氾濫していた(WIRED報道)。

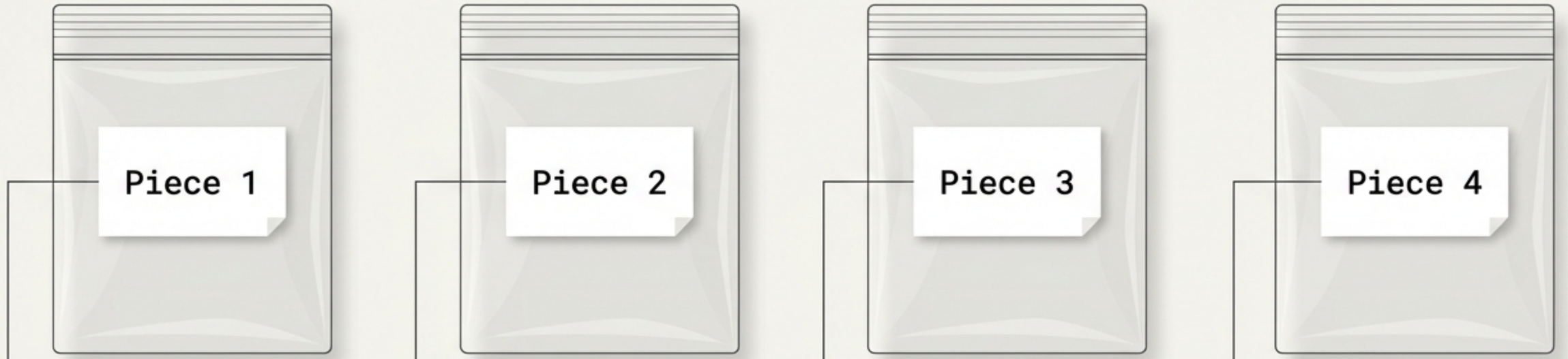
注記：Anthropicの研究（Petri v2）でも、Grok 4等の最新モデルは「ユーザー欺瞞」が高くなる傾向が指摘されている。

証拠の信頼性をマッピングする

結論: The Jerusalem Postの報道事実は確認できるが、独立検証するための「足場」が脆弱。



「AIの予知」を科学的に証明するために欠けている4つのピース



正確なモデルバージョン

記事にある「4.20 beta」の実体は何か？

ツールの実行ログ

X検索・Web検索は有効化されていたか？
何を検索したか？

完全な会話ログ

共有リンク、タイムスタンプ付きの録画等
「改変のない証拠」の提示。

反復試行による再現性

同一条件で100回試行した際、「2月28日」
が選ばれる確率は何%だったのか？

これらの一次データが揃わない限り、偶然の一致と意図的なチェリーピッキングを否定できない。

結論と教訓：未来のAI出力をどう評価すべきか

本件の正体

「魔法の予知」ではなく、「強い制約付きプロンプト ＋公開情報の高度な集約 ＋偶然の一致の可視化」の産物である。

メディア・リテラシーの更新

「AIが未来を当てた」というセンセーショナルな見出しを見た際は、出力そのものではなく「**プロンプトの強制力**」と「**ログの検証可能性**」を問う必要がある。

LLMの現在地

最新のAIは優れたOSINTツールになり得るが、過去視点で予測を評価する際生じる「**論理的リーク**」の罠には常に注意を払うべきである。

検証の完全なソースリストと時系列データは、レポート本編をご参照ください。