Evaluating Large Language Models for Enterprise Access Review Decisions

Author: Ilay Levinget

Powered by Fabrix

Abstract

Access reviews are a foundational control in enterprise identity and access management (IAM), yet they remain costly, error-prone, and difficult to scale in practice. Reviewers are required to evaluate whether access should be retained or revoked by reasoning over fragmented identity, HR, and activity data, often under time pressure and with incomplete or inconsistent context. As a result, outcomes are frequently inconsistent and difficult to defend.

In this paper, we present a real-world benchmark evaluating the ability of modern large language models (LLMs) to perform access review decisions using production-grade enterprise data. Rather than treating access reviews as a simple binary classification task, we evaluate models based on the quality, coherence, and evidentiary strength of their reasoning. Our results show that frontier LLMs can meet or exceed human-level performance, while also revealing that increased reasoning capacity does not necessarily lead to better outcomes. Finally, we describe the system-level challenges that must be addressed to make these models reliable in practice, and how Fabrix approaches those challenges.

1. Introduction

Access reviews are a core mechanism for ensuring that users retain only the permissions necessary for their role. In theory, they serve as a critical safeguard against excessive or inappropriate access. In practice, however, access reviews are widely regarded as one of the most tedious and least effective security processes in modern enterprises.

This gap between intent and outcome stems from several persistent challenges. Organizations today are highly dynamic: employees frequently change roles, teams, and responsibilities, causing access assumptions to become outdated quickly. Permissions tend to accumulate rather than expire, leading to privilege creep over time. At the same time, the data required to support informed decisions – identity attributes, HR records, group memberships, and activity signals – is often fragmented, noisy, or missing critical context. These conditions make it difficult for reviewers to confidently assess whether access remains appropriate.

The process itself further compounds the problem. Human reviewers are often asked to evaluate large volumes of similar decisions under time pressure. Fatigue, inconsistency, and reliance on heuristics are common, resulting in uneven outcomes across reviewers and review cycles.

As enterprises scale, these challenges intensify. A single organization may be required to process tens or even hundreds of thousands of access decisions each quarter. At this scale, maintaining consistency, accuracy, and defensibility becomes increasingly difficult.

Recent advances in large language models raise fundamental questions about whether access reviews can be approached differently. Can AI reason about access review decisions as effectively as, or even better than, human reviewers when faced with real-world enterprise data? And if so, which models perform best when evaluated on production-grade scenarios rather than synthetic examples?

Guiding Questions

1. Can AI reason about access review decisions as effectively as, or even better than, human reviewers when faced with real-world enterprise data?
2. If so, which models perform best when evaluated on production-grade scenarios rather than synthetic examples?

2. Benchmark Design: Grounding AI in Reality

2.1 REAL-WORLD DATASET

Rather than relying on hypothetical or simplified examples, we partnered with customers to collect real access review cases drawn directly from production environments. Each case reflects the ambiguity, inconsistency, and complexity that characterize real enterprise identity data. Together, these cases span a broad range of common access review scenarios, including:

- **Off-boarded or inactive employees**
  Accounts that should no longer retain access but remain enabled due to gaps or delays in lifecycle processes.
- **Stale or unused access**
  Permissions that exist in theory but show no meaningful activity over extended periods.
- **Over-privileged entitlements**
  Access grants that exceed what is required for a user's role or responsibilities.
- **Role and peer mismatches**
  Situations where a user's access diverges materially from comparable peers in the same role.
- **Internal transfers and temporary project access**
  Access that was appropriate during a role change or project but should now be reassessed.
- **Complex group inheritance structures**
  Permissions granted indirectly through nested groups, obscuring the original intent of access.

2.2 CONTEXT PROVIDED TO MODELS

To ensure that observed differences in performance stemmed from reasoning capability rather than privileged inputs, every model evaluated received the same information a human reviewer would typically see. This included:

- **User identity attributes and job role**
  Core identity metadata such as department, title, and employment status.
- **HR lifecycle data**
  Joiner, mover, and leaver information that frames access legitimacy over time.
- **Group and permission memberships**
  Direct and inherited access relationships across systems.
- **Peer access comparisons**
  Signals showing how a user's access compares to others in similar roles.
- **Activity and usage signals**
  Evidence of whether access is actively used or dormant.
- **Historical access review outcomes**
  Past decisions that provide continuity and precedent when available.

3. Beyond Binary Decisions: Evaluating Reasoning Quality

Access review decisions are not simple classification problems. In enterprise security, a correct decision without defensible reasoning is insufficient. Decisions must be explainable, auditable, and aligned with policy in order to be trusted and operationalized.

To capture this requirement, we introduced an AI-based evaluation framework that scores each model across multiple dimensions rather than focusing solely on outcome correctness.

3.1 EVALUATION DIMENSIONS

- **Decision correctness**
  Whether the model's keep or revoke recommendation aligns with expert-reviewed ground truth.
- **Quality of reasoning**
  Whether the model appropriately weighs relevant signals such as role alignment, peer access, and historical usage.
- **Evidence strength**
  Whether conclusions are supported by concrete, verifiable data rather than vague or generic explanations.
- **Coherence and consistency**
  Whether the reasoning is logically structured and internally consistent.

This framework rewards models that do not merely arrive at the correct answer, but that reason in a way enterprises can trust and audit.

3.2 METHODOLOGY

To evaluate model performance rigorously, we grounded the benchmark in a real-world enterprise dataset annotated by experienced IAM engineers. Each access review case in the dataset was labeled not only with a final remediation decision (retain or revoke), but also with **explicit human-authored reasoning** describing which signals were considered relevant and why.

This dual annotation approach allowed us to evaluate models along two axes: **decision alignment and reasoning fidelity**, rather than treating access reviews as a purely outcome-based task.

3.2.1 HUMAN-LABELED GROUND TRUTH

For each access review case, IAM engineers provided:

1. **Final remediation decision** – The expected keep or revoke outcome based on enterprise policy and best practice.
2. **Structured reasoning signals** – Explicit notes identifying the assets, entitlements, activity signals, peer comparisons, and lifecycle events that materially influenced the decision.

These annotations reflect how access reviews are actually performed in practice: by synthesizing multiple imperfect signals rather than relying on a single deterministic rule.

Notable Observation

This multi-layered evaluation framework allows us to distinguish between models that merely produce plausible answers and those that reason in a way that aligns with enterprise expectations, policy, and audit requirements.

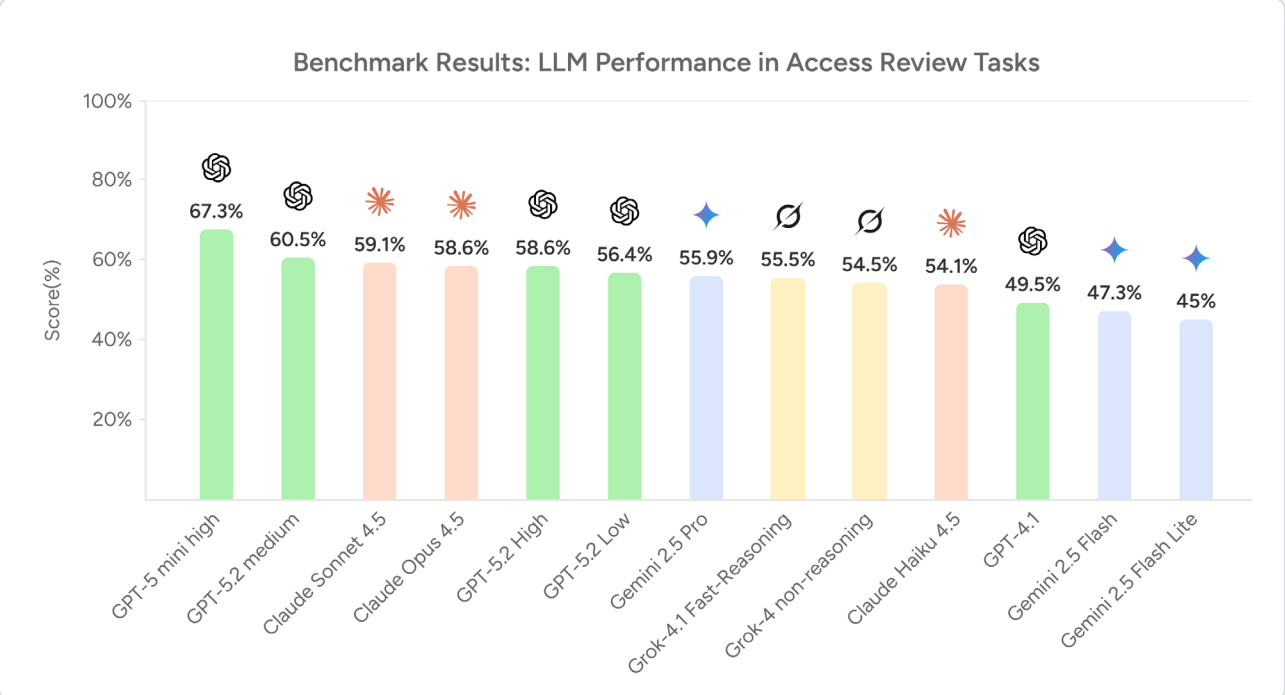
3.2.2 MODEL EVALUATION PROCEDURE

We evaluated model outputs using a combination of direct comparison and LLM-assisted judging across three dimensions:

1. **Recommendation Alignment** – We compared the model's final keep/revoke recommendation against the engineer-labeled remediation decision to assess outcome correctness.
2. **Reasoning Coverage and Signal Utilization** – Using LLM-based judges, we evaluated whether the model's explanation explicitly considered the key assets and signals identified by human reviewers as important in the original dataset. This step penalizes models that arrive at correct decisions for the wrong reasons or omit critical evidence.
3. **Reasoning Coherence and Internal Consistency** – We used LLM judges to assess whether the model's reasoning was logically structured, internally consistent, and free of contradictions or speculative leaps. This reflects the requirement that access review decisions be defensible and auditable.
4. **Remediation Alignment** – Finally, we evaluated whether the model's recommended remediation action and justification were aligned with the intent and scope of the human-labeled remediation, ensuring that the decision was not only correct in direction but appropriate in rationale.

4. Benchmark Results

Across the evaluated models, **OpenAI and Anthropic models consistently led the benchmark**, demonstrating strong performance in both decision quality and reasoning depth.



5. Key Insights from the Results

5.1 REASONING IS A FIRST-CLASS REQUIREMENT

Access reviews require synthesizing multiple weak signals rather than reacting to a single strong indicator. Models that explicitly reasoned across role expectations, peer comparisons, and activity patterns produced more defensible and auditable decisions. This reinforces a core insight: access reviews are fundamentally reasoning problems, not classification problems.

Key Takeaway #1

Access reviews are fundamentally reasoning problems, not classification problems.

5.2 WHEN MORE REASONING HURTS

One of the more unexpected findings emerged within the GPT-5.2 model family. GPT-5.2 Medium outperformed both GPT-5.2 High and GPT-5.2 Low, while the High and Low variants achieved nearly identical scores. This indicates that increased reasoning depth does not automatically improve outcomes.

In access review scenarios, excessive reasoning can introduce failure modes such as overweighting marginal or outdated signals, inventing speculative justifications, or diluting strong conclusions with unnecessary caveats. The mid-tier reasoning model struck a more effective balance by considering relevant evidence while remaining focused on the strongest signals.

Key Takeaway #2

In access review scenarios, excessive reasoning can introduce failure modes such as overweighting marginal or outdated signals, inventing speculative justifications, or diluting strong conclusions with unnecessary caveats.

5.3 SMALL GAPS MATTER AT ENTERPRISE SCALE

Although top-performing models clustered closely, even modest differences are significant in practice. In organizations processing tens or hundreds of thousands of access decisions per quarter, a 10% quality gap can result in thousands of unnecessary approvals, missed revocations of risky access, and increased manual escalations. At scale, these differences translate directly into security risk and operational cost.

Key Takeaway #3

10% quality gap can result in thousands of unnecessary approvals, missed revocations of risky access, and increased manual escalations.

6. Implications and System-Level Requirements for Access Reviews

Our findings indicate that AI systems can already perform access review decisions at or above human quality when evaluated appropriately. This demonstrates that the use of AI for access reviews is no longer experimental in nature.

However, successful adoption depends on more than model selection alone. Enterprises require access review systems that consistently produce defensible decisions, clear and evidence-backed reasoning, consistent outcomes across time and reviewers, and full auditability. Automation by itself is insufficient to meet these requirements.

At the same time, enterprise identity data presents a set of practical challenges that directly impact model performance. Permissions are often high in cardinality and heavily overlapping, activity signals may be noisy or incomplete, and organizational roles and structures change frequently. In addition, temporal inconsistencies across systems can obscure when and why access was granted or used.

Without careful system design that accounts for these realities, even strong models degrade in performance when applied in production environments. Reliable access review outcomes therefore depend not only on the capabilities of the underlying models, but also on the systems that provide context, structure, and constraints around their reasoning.

Notable Observation

Choosing a strong model is only one component of a reliable access review system. Enterprise identity data presents unique challenges:

- High cardinality and overlap of permissions
- Noisy or incomplete activity signals
- Rapidly changing roles and organizational structures
- Temporal inconsistencies across systems

7. Operationalizing Access Review Reasoning

The considerations outlined in this section reflect the system-level approaches required to make AI-driven access reviews reliable in practice. These principles are not theoretical; they are implemented in Fabrix as part of its access review system design. The goal is to translate capable model reasoning into consistent, auditable access review decisions under real enterprise conditions.

7.1 USE-CASE-SPECIFIC CONTEXT SELECTION

Not all access reviews are alike. Fabrix dynamically selects and prioritizes context based on the specific access review scenario, ensuring that models focus only on the signals that matter for that decision. This reduces noise and improves consistency.

7.2 LIFECYCLE-AWARE IDENTITY REASONING

Identity data is inherently temporal. Access may be granted temporarily, inherited indirectly, or legitimate at one point and risky later. Fabrix explicitly tracks and reasons over the lifecycle of access, incorporating grant timing, usage history, and evolution across review cycles. This temporal awareness helps avoid both false positives and missed risks.

7.3 FROM MODELS TO TRUSTED DECISIONS

Fabrix combines intelligent summarization, context filtering and normalization, noise reduction, lifecycle-aware reasoning, and alignment with organizational policies. The result is access review AI that is accurate, auditable, scalable, and policy-aligned—transforming capable models into systems enterprises can trust.

8. Conclusion

Large language models are no longer experimental tools for access reviews; they are viable decision-makers when paired with the right systems and evaluation frameworks. This benchmark demonstrates that reasoning quality matters as much as correctness, that more reasoning is not always better, that small model differences compound at enterprise scale, and that system design is as critical as model choice. At Fabrix, we will continue evaluating models against real-world enterprise data, because that is where access reviews actually happen.