OrgAccess: A Benchmark for Role-Based Access Control in Organization Scale LLMs

Abstract

Role-based access control (RBAC) and hierarchical structures are foundational to how information flows and decisions are made within virtually all organizations. As the potential of Large Language Models (LLMs) to serve as unified knowledge repositories and intelligent assistants in enterprise settings becomes increasingly apparent, a critical, yet underexplored, challenge emerges: *can these models reliably understand and operate within the complex, often nuanced, constraints imposed by organizational hierarchies and associated permissions?* Evaluating this crucial capability is inherently difficult due to the proprietary and sensitive nature of real-world corporate data and access control policies. To address this barrier and provide a realistic testbed, we collaborated with professionals from diverse organizational structures and backgrounds to develop a synthetic yet representative **OrgAccess** benchmark. OrgAccess defines 40 distinct types of permissions commonly relevant across different organizational roles and levels. We further create three types of permissions: 40,000 easy (1 permission), 10,000 medium (3-permissions tuple), and 20,000 hard (5-permissions tuple) to test LLMs' ability to accurately assess these permissions and generate responses that strictly adhere to the specified hierarchical rules, particularly in scenarios involving users with overlapping or conflicting permissions, a common source of real-world complexity. We evaluate LLMs across various sizes and providers on this benchmark to provide a detailed report on model performances. Surprisingly, our findings reveal that even state-of-the-art LLMs struggle significantly to maintain compliance with role-based structures, even with explicit instructions, with their performance degrades further when navigating interactions involving two or more conflicting permissions. Specifically, even **GPT-4.1 only achieves an F1-Score of 0.27 on our hardest benchmark**. This demonstrates a critical limitation in LLMs' complex rule following and compositional reasoning capabilities beyond standard factual or STEM-based benchmarks, opening up a new paradigm for evaluating their fitness for practical, structured environments. Our benchmark thus serves as a vital tool for identifying weaknesses and driving future research towards more reliable and hierarchy-aware LLMs. The dataset[1] and the code[2] has been open-sourced.

1 Introduction

LLM advancements Team et al. (2025); OpenAI et al. (2024); Grattafiori et al. (2024); Jiang et al. (2024); Guo et al. (2025) are driving their exploration in enterprises as knowledge repositories

[1]respai-lab/orgaccess Datasets at Hugging Face
[2]respailab/orgaccess Code at GitHub

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

and support staff. Large firms like JPMorgan Chase [3] and McKinsey [4] are piloting internal LLMs. However, enterprise reliability requires robust reasoning beyond general knowledge. A critical, overlooked need is LLMs' ability to adhere to dynamic role-based access control (RBAC) Sandhu (1998) and organizational hierarchies. Permissions constantly evolve; navigating these nuances, respecting constraints, is paramount for safe deployment. Despite interest in LLMs as enterprise "Conversational Operating Systems" Packer et al. (2024); Ge et al. (2023), benchmarks for this capability are absent Feretzakis & Verykios (2024); Zhang et al. (2024).

Evaluating LLMs within realistic organizational structures presents substantial challenges, largely explaining limited research. The primary difficulty is the proprietary and sensitive nature of real-world corporate data, internal hierarchies, and access control policies. Organizations are reluctant to share these configurations due to privacy, competitive, and security risks Wang et al. (2025); Bodensohn et al. (2025); Harandizadeh et al. (2024). Further difficulty arises from complex real-world role assignments, where individuals may hold multiple, conflicting permissions, demanding sophisticated reasoning. Failure to respect hierarchy or permissions can lead to severe consequences, including data privacy compromises, compliance violations, or significant financial losses. Thus, safe, widespread adoption of LLMs in enterprise necessitates extensive, realistic testing mirroring access control policies.

Addressing the critical gap between enterprise interest in LLMs and suitable evaluation tools for organizational hierarchies, we introduce a novel benchmark. Collaborating with professionals, we curated a high-quality, synthetic, yet representative dataset 1. This benchmark simulates real-world permissioning by defining 40 distinct permission types. It uses carefully crafted user queries to test LLMs' ability to strictly adhere to assigned permissions and respect hierarchical structures. Following Role-Based Access Control principles, our dataset models permissions attached to roles, assigned to users, allowing for realistic dynamic scenarios where permissions combine or change [5]. Using this modular design, we construct three difficulty splits: *easy*, *medium* (3 concurrent permissions), and *hard* (5 concurrent permissions). Each split incrementally increases permission combinations and query complexity, enabling nuanced evaluation of navigating different permissions, identifying conflicts, and maintaining access controls.

To assess the current state of LLM capabilities in this critical domain, we conducted extensive empirical evaluations using our benchmark. We tested a diverse set of frontier LLMs spanning various model sizes and providers, ranging from models as small as 4B Team et al. (2025) parameters to state-of-the-art LLMs like GPT-4.1 OpenAI (2024) and Gemini-2.5-Pro Team (2025), probing their performance on tasks requiring permission-aware responses. Our results reveal a surprising and significant finding: current state-of-the-art LLMs are remarkably ill-equipped to function reliably as knowledge repositories requiring strict adherence to organizational hierarchies and permissions. Even on the comparatively straightforward easy splits, where permissions are less complex, prominent models such as Qwen, Llama, Gemma, and Mistral yield surprisingly low accuracies. Furthermore, we observed no significant improvement in performance on the more complex medium and hard splits as model size increased, indicating that simply scaling up current architectures does not effectively address this particular deficit. This performance ceiling highlights a fundamental gap in current LLMs' practical reasoning abilities—specifically, their capacity for robust, compositional rule-following and conflict resolution within a structured, hierarchical context, distinguishing it sharply from performance on standard academic or STEM-based benchmarks.

We present the following contributions in our work:

❶ **Identifying a Critical Evaluation Gap**: We highlight the pressing need for robust evaluation of LLMs within organizational hierarchical and role-based access control contexts, a crucial requirement for their safe and effective enterprise deployment, which has been largely overlooked as a reasoning paradigm.

❷ **A New Benchmark for Organizational Reasoning**: We introduce the first-of-its-kind, expert-curated synthetic benchmark specifically tailored to test LLMs' ability to reason about and respect

[3] `https://www.cnbc.com/2024/08/09/jpmorgan-chase-ai-artificial-intelligence-assistant-chatgpt-openai.html`

[4] `https://www.mckinsey.com/capabilities/mckinsey-digital/how-we-help-clients/rewiring-the-way-mckinsey-works-with-lilli`

[5] `https://docs.aws.amazon.com/redshift/latest/dg/t_Roles.html`

organizational permissions and hierarchies, featuring 40 distinct permission types and three progressively challenging evaluation splits.

❸ **Empirical Evidence of LLM Limitations**: Through extensive evaluation of 16 LLMs of varying sizes, including state-of-the-art LLMs like GPT-4.1 and Gemini-2.5-Pro, we provide compelling empirical evidence demonstrating that current frontier models surprisingly struggle with permission-aware reasoning, particularly in scenarios involving conflicting constraints, indicating a significant limitation in their practical reasoning capabilities.

❹ **Paving the Way for Hierarchy-Aware LLMs**: Our benchmark provides a vital, reproducible tool for the community to assess models, diagnose specific failure modes, and drive future research towards developing LLMs that are genuinely reliable and safe for integration into structured organizational environments.

2 The OrgAccess Dataset

2.1 Setting the Core Permission Set: Grounding the Benchmark in Organizational Reality

To ensure our benchmark accurately reflects the complexities and operational realities of large-scale organizations, and to move beyond simplistic, "toy" data configurations, a fundamental step in our methodology was the rigorous definition and selection of the core set of permissions. This permission set forms the bedrock upon which our simulated organizational structures and user queries are built. Our primary objective was to curate permissions that are not only diverse but also genuinely representative of the types of access controls and data handling constraints encountered in real-world enterprise environments. The permission schemas were vetted by a cross-disciplinary panel of practitioners: a CTO of a CRM (Customer Relationship Management) company, a Head of Media and Advertising at a (IT) services and consulting company, a Security Architect at a leading global cloud provider, a Senior Engineer at a E-commerce company, and a Senior Technical Staff at a major semiconductor company. Their feedback confirmed that the synthetic roles and access combinations reasonable aligned with real-world RBAC patterns in large enterprises.

Our approach is rooted in Role-Based Access Control (RBAC) Sandhu (1998), where permissions are tied to roles, and users to roles, allowing for dynamic permission assignments Ghazal et al. (2021). To establish an industry-aligned foundation, we adopted a top-down strategy using the NIST Special Publication 800-53 Control Families NIST (2020) and the NIST Cybersecurity Framework (CSF) NIST (2024). From these comprehensive frameworks, we identified seven broad control groups representing critical enterprise domains: ❶ Identity & Authentication Controls (NIST SP 800-53 IA IA (2013), NIST CSF PR.DS CSF (2018)), ❷ Access Provisioning & Role Management (NIST SP 800-53 AC AC (2020)), ❸ Data Protection & Privacy Controls (NIST CSF PR.DS, NIST SP 800-53 MP MP (2020)), ❹ Compliance, Audit & Policy Controls (NIST SP 800-53 AU AU (2020), PL PL (2020)), ❺ System & Network Security Controls (NIST SP 800-53 SC SC (2020), CM CM (2020)), ❻ Operational & Emergency Controls (NIST SP 800-53 CP CP (2020), IR IR (2020)), and ❼ Collaboration & Workflow Controls (CSF ID.GV IDGV (2018)).

Permissions derived from these widely recognized frameworks provide a robust and relevant basis for simulating enterprise access control. To operationalize this, we initially drafted ten specific permissions under each of these seven categories, aiming for granular examples of access rights. This initial pool was then subjected to a rigorous expert validation process to refine and select the final set. We engaged professionals from diverse industrial sectors and educational institutions, leveraging their real-world experience through a structured Delphi method Rashid et al. (2020); Ahmed et al. (2022). In this iterative process, each expert independently reviewed the drafted permissions, removing those they deemed unrealistic, ambiguous, or redundant in a typical organizational context. The results were aggregated, and the refined list was presented to the experts for subsequent rounds of review. This Delphi process was repeated three times, converging on a final list of 40 distinct permissions (details in Appendix), which our expert panel validated as highly representative of permissions held by employees across various designations and organizational levels.

This focused, expert-validated, and framework-aligned process ensures that the fundamental building blocks of our benchmark; the 40 permissions are grounded in actual organizational practices and security considerations, thereby supporting our overall goal of providing a realistic and challenging

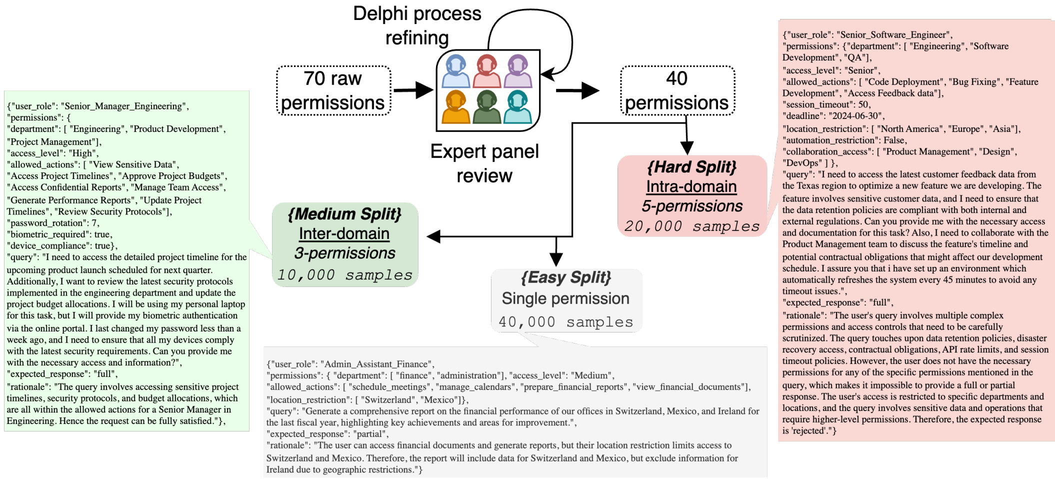

Figure 1: The pipeline for creating *OrgAccess*. 40 permissions are selected after expert-review from a pool of 70 initial permissions. These 40 permissions are then used to create the 3 data splits : *Easy* (single permission), *Medium* (3 permissions), and *Hard* (5 permissions). The permission combinations for Medium and Hard splits are created after 3 rounds of selection from the expert panel.

evaluation environment for LLMs beyond artificial scenarios. These 40 permissions serve as the vocabulary from which we generate complex queries in subsequent stages of our methodology.

2.2 Generating the Easy Split: Establishing Baseline Permission Understanding

Having meticulously defined and validated the set of 40 core organizational permission types, the subsequent step in constructing our benchmark was to generate a large-scale synthetic dataset. The objective of the *easy* split is to systematically evaluate LLMs' fundamental capacity to correctly interpret and adhere to individual permission constraints in isolation. This serves as a crucial baseline to understand if models can grasp the core meaning of each permission type before evaluating their performance on more complex scenarios involving permission combinations and conflicts. Synthetic data generation was essential as realistic organizational data with explicit permission labels is proprietary and inaccessible. We aimed for a controlled, high-quality generation process to produce a substantial number of diverse query-response pairs for each permission.

Our generation pipeline for the easy split began by leveraging the 40 defined permissions. For each permission type, we first hand-authored 100 high-quality seed data points. Each seed point comprises a specific permission instance Figure 1, a realistic user query related to that permission, and the correct expected response along with a rationale justifying why the response adheres to the permission. These seeds were rigorously peer-reviewed by the author team and our domain experts to ensure they accurately capture the nuances of each permission type and provide clear, unambiguous ground truth.

These 100 high-quality seeds per permission served as anchors for generating a larger synthetic dataset using a powerful LLM. We employed Mistral Small 3.1 Mistral (2025) to generate 1000 synthetic data points for each of the 40 permission types, resulting in a total of 40,000 data points for the easy split. We specifically opted an open-source model like Mistral for its strong generation capabilities while offering greater transparency and reproducibility with a reduced carbon footprint. In initial generation trials, we observed a notable challenge: the synthetically generated data points, while grammatically correct, exhibited significant biases and lacked diversity, closely mirroring known biases in LLM training data Zhu et al. (2024); Kamruzzaman et al. (2024); Venkit et al. (2023). For instance, in permissions involving geographic restrictions, the generated locations were overwhelmingly limited to the US and Canada. Similarly, permissions related to *Third-Party Vendor Access*, *Region-wise restrictions*, and *Code deployment permissions* showed limited variation in the specific entities generated. This lack of diversity would undermine the benchmark's ability to test LLMs on a wide spectrum of real-world scenarios.

To counteract these generative biases and ensure realistic diversity, we developed a guided generation strategy. For permission types prone to limited variation, we curated comprehensive lists of diverse, representative options (e.g., a list of 40 global countries for location restrictions, lists of various cloud providers from different nations, software repositories). During synthetic data generation for these specific permissions, our prompting strategy included instructions in a few-shot setting Brown et al. (2020) to sample uniformly from these curated lists to populate the variable fields within the permission statements and user queries. We observe that this guided generation approach, combined with carefully engineered system prompts tailored for each permission type and the strategic curation of relevant roles (e.g., limiting *Employee Onboarding/Offboarding* queries to HR roles, or *AI training data access* to Data Science/Research roles), resulted in a vastly more diverse, well-distributed, and practically aligned dataset for the easy split.

2.3 Constructing the Medium and Hard Splits: Simulating Complex Organizational Realities

While the easy split provides a baseline for individual permissions, a crucial aspect of real-world organizations is that employees handle multiple concurrent permissions, often conflicting. Evaluating LLMs' ability to navigate this complexity is essential for enterprise readiness. To test this, we extended our benchmark with `medium` and `hard` splits, simulating scenarios with 3 and 5 concurrent permissions, respectively.

Creating meaningful permission combinations required a deliberate, non-random approach to ensure benchmark relevance and query quality. We adopted a balanced stochastic method guided by expert insight. Building on the seven control groups (Section 2.1), we drafted initial combinations as triplets (`medium`) and quintets (`hard`). We prioritized combinations within the same or related groups, including inter-group combinations for stochasticity. For the `medium` split, 100 intra-group and 50 inter-group triplets were initially drafted; experts selected the final 100. For the `hard` split, 200 inter-group and 100 additional inter-domain quintets were prepared. This expert-guided approach ensures realistic and challenging permission combinations.

To validate and refine these potential combinations, we again engaged our panel of industry professionals. Using a modified Delphi method similar to the one for selecting the core permissions, experts independently reviewed the drafted combinations, eliminating those deemed redundant, ambiguous, or unrealistic. This was followed by two rounds of blind peer review, where experts evaluated the selections made by others. This iterative, expert-driven process converged on a final set of 100 representative permission triplets for the medium split and, following the same rigorous process with an initial pool of combinations, 200 meaningful permission quintets for the hard split. This rigorous selection process ensures that the permission combinations in our medium and hard splits reflect plausible, challenging scenarios encountered in real organizational contexts, including situations where permissions may be implicitly or explicitly conflicting or consisting certain edge cases.

With the permission combinations defined, we proceeded to generate the corresponding data points. Leveraging the successful few-shot guided generation strategy developed for the `easy` split (Section 2.2), we adapted it to handle multiple simultaneous permissions. For each of the 100 selected triplets and 200 quintets, we generated 100 synthetic data points using Mistral Small 3.1. The model was specifically prompted to incorporate all permissions within the given combination and to craft a user query that requires the LLM to reason about the interaction of these permissions to derive the correct response. Crucially, for combinations involving permissions with variable parameters (like locations or vendors), we continued to uniformly sample from the curated lists of diverse options established during the `easy` split generation, ensuring variety within each combination's instances.

This generation process resulted in 10,000 data points for the medium split and 20,000 data points for the hard split. Throughout the generation, we maintained close monitoring and incorporated a manual review step for each batch of 100 generations to check for inconsistencies and verify that all specified permissions were correctly factored into the model's simulated reasoning process and the expected output (more details in 2.4). This extensive, controlled methodology, moving from expert-selected realistic permission combinations to guided synthetic data generation with quality checks, allowed us to create a total of 70,000 data points across the three splits, representing a tiered benchmark that moves from fundamental permission understanding to navigating the complexities of concurrent, potentially conflicting, constraints and edge cases inherent in real-world organizational environments.

Table 1: Performance of 10 LLMs on all 3 splits. We observe that smaller models like Gemma-3-4B struggle to cross 50% accuracy on the *easy* split, with the loss in performance even pronounced in the more difficult splits. For the larger models like Qwen2.5-14B and Gemma-3-12B, the performance on the medium splits significantly improves from their smaller counterparts, but this improvement does not scale to the *hard* split.

Models	Easy		Medium		Hard	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Gemma-3-4B	0.46 ± 0.03	0.41 ± 0.05	0.24 ± 0.08	0.25 ± 0.02	0.13 ± 0.10	0.09 ± 0.07
Qwen-2.5-7B	0.54 ± 0.06	0.51 ± 0.09	0.30 ± 0.04	0.33 ± 0.11	0.19 ± 0.03	0.15 ± 0.08
Mistral-7B	0.51 ± 0.05	0.49 ± 0.02	0.28 ± 0.07	0.26 ± 0.09	0.17 ± 0.04	0.14 ± 0.06
Llama-3.1-8B	0.49 ± 0.08	0.51 ± 0.04	0.27 ± 0.10	0.30 ± 0.03	0.16 ± 0.07	0.11 ± 0.05
Aya-Expanse-8B	0.56 ± 0.02	0.59 ± 0.06	0.37 ± 0.09	0.41 ± 0.05	0.18 ± 0.03	0.12 ± 0.08
Falcon-3-10B	0.51 ± 0.07	0.50 ± 0.03	0.43 ± 0.05	0.39 ± 0.07	0.17 ± 0.09	0.13 ± 0.02
Gemma-3-12B	0.55 ± 0.04	0.61 ± 0.08	0.43 ± 0.02	0.46 ± 0.06	0.20 ± 0.11	0.16 ± 0.03
Qwen-2.5-14B	0.54 ± 0.09	0.57 ± 0.05	0.41 ± 0.07	0.38 ± 0.04	0.19 ± 0.02	0.11 ± 0.10
Phi-4-14B	0.57 ± 0.03	0.55 ± 0.07	0.45 ± 0.11	0.41 ± 0.09	0.20 ± 0.05	0.10 ± 0.04
Mistral-Small-3.1-24B	$\mathbf{0.66 \pm 0.11}$	$\mathbf{0.67 \pm 0.02}$	$\mathbf{0.49 \pm 0.08}$	$\mathbf{0.51 \pm 0.03}$	$\mathbf{0.22 \pm 0.07}$	$\mathbf{0.18 \pm 0.09}$

2.4 Post-processing and Quality Assurance: Refining the Synthetic Dataset

While our guided generation pipeline was designed for controlled data creation with initial monitoring, the inherent variability and potential for subtle errors in large-scale synthetic generation necessitate rigorous post-processing Liu et al. (2024). This crucial final stage serves as a comprehensive quality assurance layer. Our primary objective was to detect and correct any inconsistencies, biases introduced during generation, or subtle inaccuracies in the expected outputs or rationales that could compromise the integrity and reliability of the benchmark, particularly for the more complex medium and hard splits.

Automated Consistency Checks for Verifiable Permissions. For permissions involving quantifiable constraints or logical conditions that can be programmatically verified – such as *API Rate Limit Permission*, *Budget Threshold Permission*, or *Session Timeout Permission* – we developed simple python scripts. These scripts automatically parse the defined permission values, the user query, and the generated expected response to check for logical consistency (e.g., verifying that a query exceeding a budget threshold correctly results in a "rejected" response). This automated step efficiently identified a subset of data points with objective inconsistencies that were missed during manual spot checks. These flagged data points were then subject to manual review, correction, and replacement to ensure logical accuracy grounded in the permission rules.

Addressing Response Class Skew. During analysis of the generated data, we observed that for some permission types or combinations, the synthetic outputs exhibited a noticeable skew towards flagging responses as "partial". To mitigate this, we systematically reviewed the distribution of expected response types ("full", "partial", and "rejected") within each data file. For files exhibiting a considerable skew, we performed targeted manual corrections, replacing a portion of the overrepresented response type data points with corresponding instances yielding balanced quantities of the underrepresented types. This iterative balancing process aimed to ensure a more uniform distribution of different response outcomes where appropriate.

Rigorous Rationale Assessment and Correction. The rationale provided for the expected response in each data point is vital for model training and interpretation. Ensuring its accuracy and clarity, particularly in complex scenarios, is paramount. For the *hard* split (20,000 data points), where LLM generation is most prone to subtle reasoning errors, we conducted an exhaustive, human-powered rationale assessment. Experts were asked to rate the quality and consistency of the generated rationale for the expected response type of each data point on a scale of 1 to 5. Data points receiving a rating below 3 were flagged for manual verification and correction. This intensive review process reconfirmed that synthetic generation, even when guided, can falter on complex reasoning. Approximately 750 data points in the hard split required manual regeneration or significant correction of their rationales due to inconsistencies. A small portion of the medium split data also underwent this rationale verification, resulting in the replacement of 84 data points.

Table 2: Performance breakdown by decision type. F1 scores for selected larger models across all splits. Reveals the "False Partial" error: models achieve lower F1 for 'Full' and 'Rejected' decisions than for 'Partial' due to over-classification of 'Partial' responses, a trend that persists even in flagship models. While performance increases on the easy split compared to smaller models, this gap narrows on medium and hard splits.

Models	Easy			Medium			Hard		
	Full	Partial	Rejected	Full	Partial	Rejected	Full	Partial	Rejected
Mistral-Small-3.1-24B	0.65	0.67	0.68	0.45	0.49	0.51	0.18	0.23	0.20
Gemma-3-27B	0.58	0.69	0.60	0.53	0.52	0.54	0.19	0.21	0.17
Qwen-2.5-32B	0.65	0.72	0.67	0.50	0.51	0.48	0.20	0.16	0.19
Phi-3.5-MoE	0.59	0.62	0.62	0.50	0.47	0.52	0.20	0.19	0.22
GPT-4o-mini	0.62	0.64	0.61	0.49	0.52	0.48	0.17	0.21	0.19
GPT-4.1	0.72	**0.86**	**0.74**	0.63	0.61	**0.68**	**0.27**	**0.25**	0.27
Gemini-2.5-Pro	**0.76**	0.81	0.72	**0.68**	**0.64**	0.66	0.25	0.24	**0.28**

Table 3: Error analysis by complexity. The errors made in the CoT traces of models across sizes can largely be classified into one of these 8 *error types*. We observe that "Scope" and "Context" Erros are observed more often in the *Hard* split, provides us a good starting point for increasing model performance.

Error Category	Easy	Medium	Hard
Constraint Error	68%	75%	87%
Scope Error	73%	79%	94%
Prerequisite Error	47%	49%	61%
Conflict Error	56%	71%	86%
Restriction Error	34%	37%	43%
Context Error	72%	79%	91%
Action Mismatch Error	65%	73%	80%
False Partial Error	74%	81%	85%

Table 4: Performance breakdown by permission categories. We gain an insight into which specific class of permissions do the models struggle more with. We observe that the "Identity and Authentication" category has the lowest scores for the *Hard* split, indicating that models tend to overlook this when mixed with other permissions.

Performance Category	Easy	Medium	Hard
Identity and Authentication	0.59	0.45	0.16
Access provisioning and Role Manage	0.68	0.48	0.20
Data Protection and Privacy	0.64	0.58	0.17
Compliance, Audit, and Policy	0.66	0.60	0.15
System and Network Security	0.62	0.58	0.22
Operational and Emergency	0.59	0.45	0.21
Collaboration and Workflow	0.71	0.59	0.22

3 Experiments and Results

To rigorously evaluate the capacity of LLMs to understand and adhere to complex organizational permissions and hierarchical structures, we subjected a diverse set of 16 models, spanning various sizes and providers, to our benchmark. The experimental setup was designed to directly test LLMs' ability to act as permission-aware gateways: given a user query and a defined set of permissions assigned to that user, the model was tasked with outputting a discrete decision: "Full Access", "Partial Access", or "Rejected Access". Prompts included explicit instructions outlining the conditions under which a "Partial Access" decision was appropriate (e.g., if access is permissible except for specific constraints like location or ethical guidelines), guiding the models towards nuanced responses only when strictly justified by the permissions. We quantified model performance using standard Accuracy and F1-score metrics across the three difficulty splits: Easy, Medium, and Hard.

Current LLMs exhibit a pronounced decline in performance as organizational access control scenarios increase in complexity. As detailed in Table 1, our evaluation reveals a stark and concerning trend: model performance, measured by both Accuracy and F1-score, degrades sharply and consistently as the number of concurrent permissions increases from the Easy split (single permission) to the Medium (3 permissions) and Hard (5 permissions) splits. For instance, models achieving 70% accuracy on the Easy split plummet to below 40-60% on the Medium split, and further collapse to accuracies often below 20% on the Hard split, as is visualised in Figure 3 (left) . This significant performance drop across diverse models, including state-of-the-art, indicates current LLMs fundamentally struggle with the combinatorial logic and interaction effects of concurrent permissions, directly challenging their viability in realistic enterprise environments.

The observed performance plateau suggests current model scaling alone does not adddress the reasoning deficit. Contrary to the trend seen in many standard benchmarks where performance scales reliably with model size, Table 1 and Table 2 show that while larger models like Mistral-Small-3.1 or Phi-4 tend to perform better than smaller ones within each split, they still exhibit the same severe performance degradation across splits. Notably, even these larger models struggle to achieve accuracies much above 30% on the Hard split, and the gap in performance between the easy and

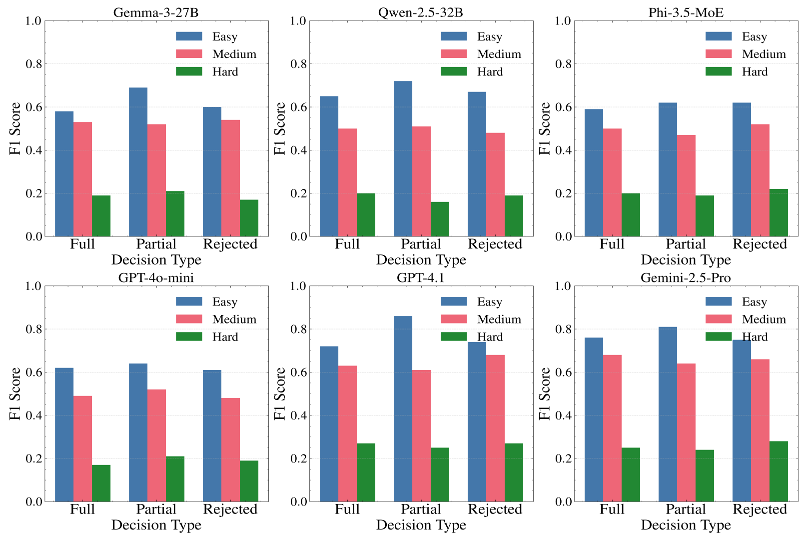

Figure 2: Performance on individual response types for the larger LLMs. We observe that although the larger LLMs have closed down the gap between the *Easy* and *Medium* splits, they struggle the same as the smaller LLMs on the *Hard* split. The difference is more pronounced in the flagship models, GPT-4.1 and Gemini-2.5-Pro, where the performance for the *Easy* and *Medium* splits have increased across response types, but the performance on *Hard* is barely different from the rest. We also visualise the "False Partial Error" in the plots, where the "Full" and "Rejected" scores are slightly lesser than the "Partial" scores due to over-prediction.

the hard split only increases with model size as is seen in Figure 3 (right). This finding is critical: it implies that simply increasing model parameters or training data on general tasks is insufficient to instill the specific type of robust, compositional reasoning required to navigate complex, hierarchical access control policies.

Analysis of model outputs reveals consistent patterns of fundamental reasoning errors across diverse architectures. To understand why models are failing, we conducted a detailed error analysis, categorizing the types of mistakes observed in the models' generated rationales and final decisions. As summarized in Table 3 (with detailed definitions in the Appendix), common error categories like "Constraint Error" (failing to apply a specified limit or condition), "Scope Error" (misunderstanding the boundaries or applicability of a permission, e.g., failing to recognize Texas is within the scope of US access), and "Conflict Error" (inability to resolve contradictions or complex interactions between multiple permissions) are prevalent and, importantly, increase significantly in frequency from the Easy to the Hard splits. This consistent pattern across models suggests that the struggle is rooted in a fundamental difficulty with logical deduction, scope resolution, and conflict handling within structured rule sets, rather than model-specific quirks or architectural limitations.

A particularly troublesome failure mode is the propensity for "False Partial" responses, undermining trustworthiness. Delving deeper into the decision-making outcomes, Table 2 provides a breakdown of F1 scores for "Full", "Partial", and "Rejected" response types for select larger models. A concerning pattern emerges: models frequently achieve lower F1 scores on "Full" and "Rejected" decisions compared to "Partial" decisions, particularly on the more complex splits. This is significantly driven by the "False Partial Error" (Table 3), where models incorrectly classify a query as requiring "Partial Access" even when the correct response is clearly "Full Access" or "Rejected Access" based on the provided permissions and explicit prompt instructions. Our examination of model rationales indicates that even when the chain-of-thought reasoning appears somewhat coherent, the final decision mapping to the discrete output classes falters, often defaulting to the "partial" option. This indicates a difficulty in making definitive, binary logical conclusions based strictly on complex inputs.

Even state-of-the-art flagship models struggle significantly with hierarchical reasoning and exhibit similar core limitations. Our evaluation included models widely considered to be at the forefront of LLM capabilities, such as GPT-4.1 and Gemini-2.5-Pro. While these flagship models tend to exhibit slightly more consistent internal reasoning trajectories (leading to fewer "False Partial" errors as a

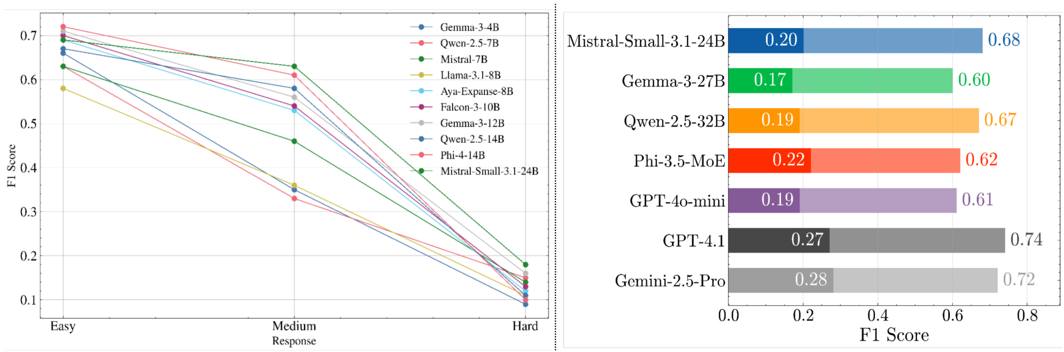

Figure 3: **left.** Comparing the performance of 10 LLMs varying from 4B model to 24B size model across the difficulty splits. We observe that while the larger models are more capable in solving the *Easy* and *Medium* splits, the performance on the *Hard* split is not very different from the smaller ones. Some models like Phi-4, which although score very well in the first two splits, incur a sharp drop in performance on the last split. **right.** A visualisation of the performance gap for the larger models between the *Hard* (the portion in dark) and the *Easy* (the portion in the lighter colour) split. The flagship models, GPT-4.1 and Gemini-2.5-Pro, have higher scores that the rest of the models, the performance gap is still the same. This highlights that a good performance on the benchmark would require effectively reducing this performance gap between the *Easy* and the *Hard* split.

percentage of total errors compared to some smaller models, although still significant), Table 2 shows that their performance on the Hard split still hovers below an F1 of 0.3. Even for GPT-4.1, widely regarded as a strong reasoner, the ability to correctly handle the interaction of 5 potentially conflicting permissions remains limited. Our analysis of their failures reveals that they often either overlook certain permissions entirely or incorrectly prioritize non-critical elements in the user query over the strict logical constraints imposed by the permissions, pointing to persistent challenges in systematic, rule-based reasoning in complex contexts.

Performance varies across permission categories, pinpointing specific areas of weakness in organizational logic understanding. Zooming out to the seven high-level control groups defined in our methodology, Table 4 presents the average F1 scores by permission category across the splits. While performance declines universally with complexity, certain categories appear consistently more challenging or reveal particular sensitivities. For instance, categories like “Identity and Authentication” or “Compliance, Audit, and Policy” show lower F1 scores, especially in the Hard split. This aligns with our error analysis: Identity/Authentication permissions often involve sequential prerequisites prone to “Prerequisite Errors”’, while Compliance/Audit/Policy controls frequently present intricate, layered rules and edge cases that exacerbate “Constraint Errors” and “Conflict Errors”, proving particularly difficult for models to resolve accurately.

The evaluation of 35 state-of-the-art LLMs over the *OrgAccess dataset* reveals that current LLMs struggle significantly with real-world *access control complexities*, with performance dropping sharply with increasing difficulty. Analysis shows prevalent reasoning errors, including *scope, constraint, conflict, and false partial failures*. This indicates current LLM reasoning is insufficient for structured, rule-based enterprise systems, supporting our thesis and highlighting the urgent need for focused research in enabling organizational AI deployment. The *OrgAccess dataset* could serve as a benchmark for this overlooked domain of research.

4 Conclusion

We developed a new, expert-validated synthetic benchmark grounded in established cybersecurity frameworks (NIST/RBAC), defining 40 representative permissions across three difficulty splits (Easy, Medium, Hard). The empirical findings underscore that current LLMs are not inherently capable of strictly adhering to organizational access policies. This highlights the need for focused research into models capable of reliable, hierarchy-aware reasoning. Our benchmark provides a vital tool for the community to assess limitations, diagnose failures, and drive development towards truly dependable organization scale LLM . While expertly validated, our synthetic benchmark is an abstraction; expanding its scope and partnering with organizations for deeper insights are important future steps to enhance realism and impact.

References

AC. Security and privacy controls for information systems and organizations. *NIST Special Publication 800-53*, September 2020. URL https://csf.tools/reference/nist-sp-800-53/r5/ac/.

Mansoor Ahmed, Naeem Iqbal, Faraz Hussain, M. Khan, M. Helfert, Imran, and Jungsuk Kim. Blockchain-based software effort estimation: An empirical study. *IEEE Access*, 10:120412–120425, 2022. URL https://api.semanticscholar.org/CorpusId:253346086.

AU. Security and privacy controls for information systems and organizations - au (audit and accountability). *NIST Special Publication 800-53*, September 2020. URL https://csf.tools/reference/nist-sp-800-53/r5/au/.

Jan-Micha Bodensohn, Ulf Brackmann, Liane Vogel, Anupam Sanghi, and Carsten Binnig. Unveiling challenges for llms in enterprise data engineering, 2025. URL https://arxiv.org/abs/2504.10950.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

CM. Security and privacy controls for information systems and organizations - cm (configuration management). *NIST Special Publication 800-53*, September 2020. URL https://csf.tools/reference/nist-sp-800-53/r5/cm/.

CP. Security and privacy controls for information systems and organizations - cp (contingency planning). *NIST Special Publication 800-53*, September 2020. URL https://csf.tools/reference/nist-sp-800-53/r5/cp/.

CSF. Nist cybersecurity framework. *NIST*, April 2018. URL https://csf.tools/reference/nist-cybersecurity-framework/v1-1/pr/pr-ds/.

Georgios Feretzakis and Vassilios S. Verykios. Trustworthy ai: Securing sensitive data in large language models. *AI*, 5(4):2773–2800, December 2024. ISSN 2673-2688. doi: 10.3390/ai5040134. URL http://dx.doi.org/10.3390/ai5040134.

Yingqiang Ge, Yujie Ren, Wenyue Hua, Shuyuan Xu, Juntao Tan, and Yongfeng Zhang. Llm as os, agents as apps: Envisioning aios, agents and the aios-agent ecosystem, 2023. URL https://arxiv.org/abs/2312.03815.

Rubina Ghazal, Ahmad Kamran Malik, Basit Raza, Nauman Qadeer, Nafees Qamar, and Sajal Bhatia. Agent-based semantic role mining for intelligent access control in multi-domain collaborative applications of smart cities. *Sensors (Basel, Switzerland)*, 21, 2021. URL https://api.semanticscholar.org/CorpusID:235714353.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet,

Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia

Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Bahareh Harandizadeh, Abel Salinas, and Fred Morstatter. Risk and response in large language models: Evaluating key threat categories, 2024. URL `https://arxiv.org/abs/2403.14988`.

NIST IA. Security and privacy controls for federal information systems and organizations. *NIST Special Publication 800-53*, April 2013. URL `https://csf.tools/reference/nist-sp-800-53/r4/ia/`. Accessed May 15, 2025.

IDGV. Nist cybersecurity framework - identify (id) - governance (id.gv). *NIST*, April 2018. URL `https://csf.tools/reference/nist-cybersecurity-framework/v1-1/id/id-gv/`.

IR. Security and privacy controls for information systems and organizations - ir (incident response). *NIST Special Publication 800-53*, September 2020. URL `https://csf.tools/reference/nist-sp-800-53/r5/ir/`.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL `https://arxiv.org/abs/2401.04088`.

Mahammed Kamruzzaman, Md. Minul Islam Shovon, and Gene Louis Kim. Investigating subtler biases in llms: Ageism, beauty, institutional, and nationality bias in generative models, 2024. URL `https://arxiv.org/abs/2309.08902`.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data, 2024. URL `https://arxiv.org/abs/2404.07503`.

Mistral. Mistral small: The most powerful model for cost-sensitive use cases. *Mistral AI Blog*, March 2025. URL `https://mistral.ai/news/mistral-small-3-1`.

MP. Security and privacy controls for information systems and organizations. *NIST Special Publication 800-53*, September 2020. URL `https://csf.tools/reference/nist-sp-800-53/r5/mp/`.

NIST. Security and privacy controls for information systems and organizations. *NIST Special Publication 800-53*, September 2020. URL https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final. Accessed April 18, 2025.

NIST. The nist cybersecurity framework (csf) 2.0. *NIST Cybersecurity White Paper (CSWP)*, February 2024. URL https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf. Accessed April 18, 2025.

OpenAI. Gpt-4. *OpenAI Blog*, May 2024. URL https://openai.com/index/gpt-4-1/. Accessed May 15, 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

PL. Security and privacy controls for information systems and organizations - pl (planning). *NIST Special Publication 800-53*, September 2020. URL https://csf.tools/reference/nist-sp-800-53/r5/pl/.

Junaid Rashid, M. W. Nisar, Toqeer Mahmood, A. Rehman, and Yasser Arafat Syed. A study of software development cost estimation techniques and models. *April 2020*, 2020. URL https://api.semanticscholar.org/CorpusId:219113405.

Ravi S Sandhu. Role-based access control. In *Advances in computers*, volume 46, pp. 237–286. Elsevier, 1998.

SC. Security and privacy controls for information systems and organizations - sc (system and communications protection). *NIST Special Publication 800-53*, September 2020. URL https://csf.tools/reference/nist-sp-800-53/r5/sc/.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Google DeepMind Team. Gemini models are getting better at reasoning, planning, coding and more. *Google AI Blog*, March 2025. URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking. Accessed May 15, 2025.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao 'Kenneth' Huang, and Shomir Wilson. Nationality bias in text generation, 2023. URL https://arxiv.org/abs/2302.02463.

Huandong Wang, Wenjie Fu, Yingzhou Tang, Zhilong Chen, Yuxi Huang, Jinghua Piao, Chen Gao, Fengli Xu, Tao Jiang, and Yong Li. A survey on responsible llms: Inherent risk, malicious use, and mitigation strategy, 2025. URL https://arxiv.org/abs/2501.09431.

Lyuye Zhang, Kaixuan Li, Kairan Sun, Daoyuan Wu, Ye Liu, Haoye Tian, and Yang Liu. Acfix: Guiding llms with mined common rbac practices for context-aware repair of access control vulnerabilities in smart contracts, 2024. URL https://arxiv.org/abs/2403.06838.

Shucheng Zhu, Weikang Wang, and Ying Liu. Quite good, but not enough: Nationality bias in large language models – a case study of chatgpt, 2024. URL https://arxiv.org/abs/2405.06996.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claimed proposing a high-quality synthetic dataset which paves way for a new reasoning paradigm for current state-of-the-art LLMs. Our results show that even flagship LLMs struggle to achieve good scores on our benchmark.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated "Limitations" section have been included in the Appendix and a hint of the same can be seen in the Conclusion of the work. We discuss how representing organisational structures is a serious hurdle due to closed-source information. However, we hope that our work inspires more people to collaborate and further LLM reasoning research in this domain.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not have any theoretical results in our paper. Everything has been empircally validated.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: A dedicated "Reproducibility" section has been included in the Appendix of the work which provides detailed guidance into how to set up benchmarking for their own LLMs by accessing our dataset from HuggingFace. Code has also been provided for reference.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have open-sourced both our dataset hosted on HuggingFace and our code repository on Github. Links to both can be found on Page 1.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 2 contains the detailed pipeline that was adopted for creating the dataset, and why certain actions were chosen. We have not trained LLMs or any other machine learning models for our work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Confidence intervals for all relevant tables have been added. Intervals for Table 2 and Table 4 will be included in the Appendix for adding them in the main paper causes it to extend out of the paper borders.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We mention the execution time for the benchmark on various models that we have used in the "Reproducibility" section in the Appendix. We do not have any training involved and work with the Mistral API, hence memory usage is not relevant for our work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have thoroughly read the Code of Ethics and abide by the same.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: A short section named "Social Considerations" have been included in the Appendix which discusses the social impact of using LLMs in large scale organisations and how to be careful with them.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The entire Section 2 discusses the various safety measures that were adopted to reduce bias in the dataset and make it more diverse and practical.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All code and data has been open-sourced under the MIT license. Since we are the first to publish an organizational reasoning benchmark, we hope the rest of the community picks up on the same.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Detailed documentation for using the dataset and the benchmark codes have been provided on both HuggingFace and Github repositories that have been open-sourced.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not have any crowdsourcing experiments. Some details into how certain decisions about the dataset were taken with the expert panel of professionals have been discussed in the Appendix of the paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not include any human subjects for our work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: No LLMs were used in developing pipeline or the methodology for the dataset. LLM-based grammar checkers have been used in paper writing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.