DMind Benchmark: Toward a Holistic Assessment of LLM Capabilities across the Web3 Domain

Enhao Huang[*]   **Pengyu Sun**[*]   **Zixin Lin**[*]   **Alex Chen**[*†]   **Joey Ouyang**[*†]
Haobo Wang[*]   **Dong Dong**[*]   **Gang Zhao**[†]   **James Yi**[†]   **Frank Li**[†]   **Ziang Ling**[†]
Lowes Yang[†‡]

[*]Zhejiang University
[†]DMind.ai
[‡]Corresponding author: team@dmind.ai

Abstract

Large Language Models (LLMs) have achieved impressive performance in diverse natural language processing tasks, but specialized domains such as Web3 present new challenges and require more tailored evaluation. Despite the significant user base and capital flows in Web3—encompassing smart contracts, decentralized finance (DeFi), non-fungible tokens (NFTs), decentralized autonomous organizations (DAOs), on-chain governance, and novel token-economics—no comprehensive benchmark has systematically assessed LLM performance in this domain. To address this gap, we introduce the **DMind Benchmark**, a holistic Web3-oriented evaluation suite covering nine critical subfields: fundamental blockchain concepts, blockchain infrastructure, smart contract, DeFi mechanisms, DAOs, NFTs, token economics, meme concept, and security vulnerabilities. Beyond multiple-choice questions, DMind Benchmark features domain-specific tasks such as contract debugging and on-chain numeric reasoning, mirroring real-world scenarios. We evaluated 26 models—including ChatGPT, Claude, DeepSeek, Gemini, Grok, and Qwen—uncovering notable performance gaps in specialized areas like token economics and security-critical contract analysis. While some models excel in blockchain infrastructure tasks, advanced subfields remain challenging. Our benchmark dataset and evaluation pipeline are open-sourced on `https://huggingface.co/datasets/DMindAI/DMind_Benchmark`, reaching #1 in Hugging Face's trending dataset charts within a week of release.

1   Introduction

Recent advancements in Large Language Models (LLMs) have demonstrated their profound capabilities across a wide spectrum of natural language processing (NLP) tasks [38, 14, 48, 47]. With their maturation from experimental research to production-ready systems, LLMs are increasingly being deployed in specialized domains. Fields such as biomedical informatics [43], finance [57], legal analysis [31], and software engineering [10] are actively integrating these models, recognizing that deep, domain-specific knowledge is paramount for achieving reliable and impactful results.

This proliferation of domain-specific LLM applications underscores an urgent need for specialized evaluation frameworks. Prevailing benchmarks like MMLU [19], BIG-Bench [22], and HELM [27], while offering valuable insights into general linguistic competence, fall short in assessing the nuanced knowledge and sophisticated reasoning demanded by high-stakes sectors. Consequently, fields like healthcare, finance, and regulatory compliance, where errors carry significant repercussions, have
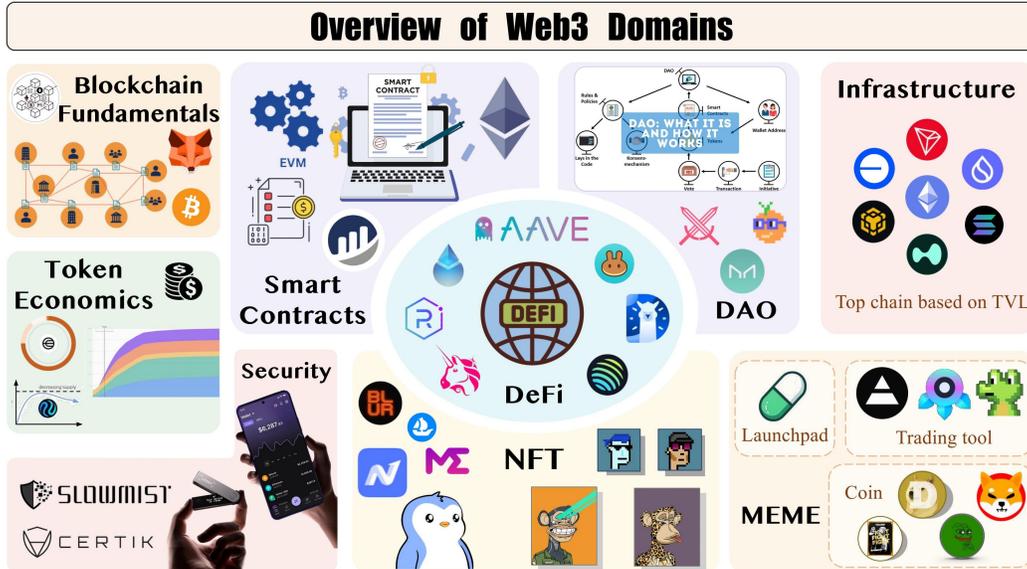
Figure 1: Illustration of the multifaceted Web3 ecosystem, visually mapping its core components into the nine key domains evaluated by the DMind Benchmark: Blockchain Fundamentals, Smart Contracts, DeFi, NFTs, DAOs, Token Economics, Security, Infrastructure, and Meme concepts. Each domain is represented by characteristic concepts and associated imagery reflecting its core attributes.

seen the development of bespoke benchmarks for rigorous expertise validation [11, 23]. Yet, to the best of out knowledge, a correspondingly comprehensive evaluation framework for the nascent and intricate domain of Web3 is conspicuously missing.

Web3, as a paradigm shift to a user-centric, decentralized internet, relies on cryptographic and distributed technologies to curtail dependence on trusted intermediaries [7, 55, 9, 15]. Its scope extends beyond blockchain protocols or Decentralized Finance (DeFi) [39], encompassing a diverse array of concepts including Non-Fungible Tokens (NFTs) [52], Decentralized Autonomous Organizations (DAOs) [5], on-chain governance, privacy-enhancing infrastructures, and innovative cryptoeconomic primitives. Navigating this multifaceted ecosystem necessitates a profound interdisciplinary understanding of cryptography, distributed systems, economics, and game theory. The swift evolution of on-chain applications, coupled with substantial financial stakes, amplifies the demand for accurate and robust AI-driven solutions. Consequently, the proficiency of LLMs within Web3 carries significant implications for user experience, security, and the broader adoption of these decentralized technologies, especially considering its large user base and considerable capital flows.

Notwithstanding the escalating significance of Web3, a comprehensive benchmark for evaluating LLM proficiency in its core tasks has been notably absent. A critical chasm persists between the rapidly innovating Web3 community—pioneering advancements in smart contracts and token economics—and the AI community, which is concurrently developing larger models and novel training paradigms at an accelerated rate. This disconnection has impeded systematic model performance assessment and the precise identification of areas requiring targeted enhancements for Web3 applications.

To bridge this critical gap, we introduce the **DMind Benchmark**, the inaugural holistic evaluation suite meticulously engineered to assess LLM performance within the Web3 domain. Our benchmark spans nine pivotal subfields: *(1) fundamental blockchain concepts, (2) blockchain infrastructure, (3) smart contract, (4) DeFi mechanisms, (5) DAOs, (6) NFTs, (7) token economics, (8) meme concepts, and (9) security vulnerabilities*. Beyond multiple-choice questions gauging foundational understanding, the DMind Benchmark incorporates a spectrum of domain-specific subjective tasks, including smart contract debugging, numerical reasoning over on-chain data, and security auditing. These tasks are designed to emulate real-world challenges, thereby offering a granular assessment of LLM capabilities under practical operational conditions.

Employing the DMind Benchmark, we conducted a rigorous evaluation of 26 prominent LLMs, including the ChatGPT [38], Claude [1], DeepSeek [14], Gemini [46], Grok [58], and Qwen [2] series, revealing significant performance disparities. While some leading models exhibited proficiency in foundational Web3 concepts, many faltered in highly specialized or rapidly advancing subfields, such as token economics and security-sensitive smart contract. Our findings indicate that distinct model families show varying strengths, yet generally display consistent performance scaling within their lineage. Notably, while models excelled in blockchain infrastructure tasks, their performance was merely moderate in areas like fundamental blockchain principles, smart contract, DeFi mechanisms, DAOs, and security vulnerabilities. Furthermore, nascent fields like token economics and meme concepts presented substantial challenges, highlighting an urgent imperative for targeted model enhancements and robust evaluations on advanced or evolving Web3 topics.

Our primary contributions are threefold: **(1)** We introduce the **DMind Benchmark**, the first comprehensive, Web3-focused evaluation framework designed to unify efforts between the AI and blockchain research communities. **(2)** We provide a rigorous assessment of an extensive set of leading LLMs, pinpointing their respective strengths and weaknesses across crucial Web3 functionalities. **(3)** We have open-sourced the DMind Benchmark dataset and its associated evaluation pipeline[1]. The benchmark's rapid ascent to the #1 position on Hugging Face's trending dataset charts within one week of its release attests to its timeliness and perceived importance by the community. We contend that the DMind Benchmark will catalyze the development of more specialized and resilient LLMs. More broadly, by establishing a rigorous evaluation framework for LLMs in the complex and rapidly evolving Web3 domain, this work provides a critical testbed that can spur further AI research into robust domain adaptation, specialized reasoning, and the development of more capable and trustworthy intelligent systems.

2 Related Work

2.1 LLM Evaluation Benchmarks

Evaluating the capabilities of Large Language Models (LLMs) has garnered significant attention, leading to numerous benchmarks assessing different facets of model performance. Early general-purpose benchmarks like GLUE [50] and SuperGLUE [51] focused primarily on natural language understanding. More recent and comprehensive efforts, including MMLU [19], BIG-Bench [22], and HELM [27], provide broader assessments of advanced capabilities such as higher-level reasoning, domain knowledge, and instruction-following proficiency. MMLU evaluates models across 57 diverse subject areas; BIG-Bench incorporates over 200 tasks designed to probe aptitudes beyond conventional NLP benchmarks; and HELM offers a framework to assess multiple dimensions like accuracy, calibration, robustness, fairness, and efficiency.

While these general benchmarks offer invaluable insights, they often do not explicitly address the specialized demands of niche domains. This limitation has spurred the creation of domain-specific benchmarks to rigorously evaluate models in specialized areas. For instance, in the medical field, MedQA [23], MultiMedQA [42], and MedMCQA [40] examine medical knowledge and diagnostic reasoning. Similarly, finance has seen benchmarks like FinBen [11] and FinEval [17] for assessing the understanding of financial concepts and analytical capabilities. Other notable examples include LegalBench [16] for legal reasoning, CyberBench [29] for cybersecurity knowledge, and SafetyBench [61] for evaluating model safety in critical scenarios. Such targeted evaluations underscore the importance of domain-specific assessment for advancing LLM performance in highly specialized settings. Despite these advancements, to the best of our knowledge, a benchmark specifically for evaluating LLM capabilities within the Web3 domain—characterized by its technical intricacies, interdisciplinary nature, and critical security considerations—has been notably absent.

2.2 Web3 Technologies and Applications

Web3 signifies a shift from centralized Web2 to a decentralized ecosystem built on blockchain technology, prioritizing user control, decentralization, and trustlessness [60, 59]. This overview consolidates Web3's critical facets into four interconnected pillars: foundational infrastructures, decentralized applications, governance and community, and security imperatives.

---

[1] https://huggingface.co/datasets/DMindAI/DMind_Benchmark

Foundational Infrastructures. Web3's core fundamentally relies on distributed ledgers, secure consensus mechanisms, and advanced cryptography for true transparency and immutability [35, 6]. Layer-1 networks and innovative Layer-2 scaling solutions collaboratively address scalability and interoperability, forming the crucial trustless foundation for decentralized applications [4, 63].

Smart Contracts and Decentralized Applications. Smart contracts automate agreements on blockchains, removing intermediaries [45, 3]; their analysis is vital for security (e.g., addressing reentrancy, overflows) and efficiency [28, 25, 41]. Key dApps include Decentralized Finance (DeFi), offering trustless financial services [12, 54], and Non-Fungible Tokens (NFTs), enabling unique digital asset ownership and fostering new economies [52, 34].

Governance and Community. Decentralized Autonomous Organizations (DAOs) enable collective governance via token voting, fostering community-led initiatives [53, 18]. Token economics (tokenomics) guide incentives and network growth via token lifecycle management [21, 8]. Meme-driven phenomena also significantly impact Web3 adoption and community building [30, 24].

Security Imperatives. Web3's decentralization mandates robust security against threats like flash loan exploits, rug pulls, and Sybil attacks [20]. Mitigation relies on rigorous audits, formal verification, and proactive vulnerability disclosures to protect assets and maintain trustless system integrity.

Together, these pillars illustrate Web3's interdisciplinary nature, bridging cryptography, distributed systems, economics, and social governance. Effective modeling and evaluation within this complex domain thus demand sophisticated language understanding coupled with the ability to synthesize advanced, interconnected concepts across technology, finance, and community practices.

2.3 LLMs for Web3 Applications

Recent studies highlight the significant strides LLMs are making in empowering the Web3 domain [32]. Notably, they are enhancing smart contract security through improved vulnerability detection [56] and accelerating development via automated code generation [37, 36, 62], while also streamlining documentation support [44, 13]. Furthermore, LLMs are offering deeper insights via sophisticated blockchain data analytics [49], aiding in cryptocurrency price forecasting [26], and enabling more intuitive DeFi protocol interactions [33], thereby catalyzing innovation and development across the Web3 ecosystem.

3 Framework of DMind Benchmark

3.1 DMind Benchmark Data Source

The development of the DMind Benchmark was rooted in a rigorous data collection and expert-driven curation process, aimed at ensuring its ecological validity and comprehensive coverage of topics pertinent to the contemporary Web3 landscape. The construction process involved several key stages:

First, an extensive reconnaissance phase surveyed 39 prominent Web3-focused communities, forums, and media outlets. This broad-spectrum investigation aimed to capture diverse information and discussions of pressing interest and concern to Web3 end-users and developers. This initial data gathering effort yielded a 5.7 GB multimodal corpus with rich content including textual articles, discussions, and relevant imagery that reflect real-world Web3 discourse.

Subsequently, this extensive raw dataset was meticulously analyzed and synthesized by five dedicated domain specialists. Each expert brought over eight years of direct experience from various sectors of the Web3 field. Their collective expertise helped distill the most salient themes, challenging concepts, frequently encountered problems, and emerging trends from the 5.7 GB of collected information. This expert panel was tasked with transforming these insights into a structured set of evaluative questions and tasks that reflect the complexities and practical challenges of the Web3 domain.

This intensive, expert-led curation culminates in the DMind Benchmark, comprising 1,917 unique questions (1283 single-choice, 586 multiple-choice, 48 subjective). These questions are carefully designed to assess a deep and nuanced understanding across the nine pivotal Web3 subfields outlined previously, moving beyond surface-level knowledge to probe critical reasoning and problem-solving abilities. This systematic approach to data sourcing and question development underpins the benchmark's capacity to provide meaningful insights into LLM performance in the Web3 domain.
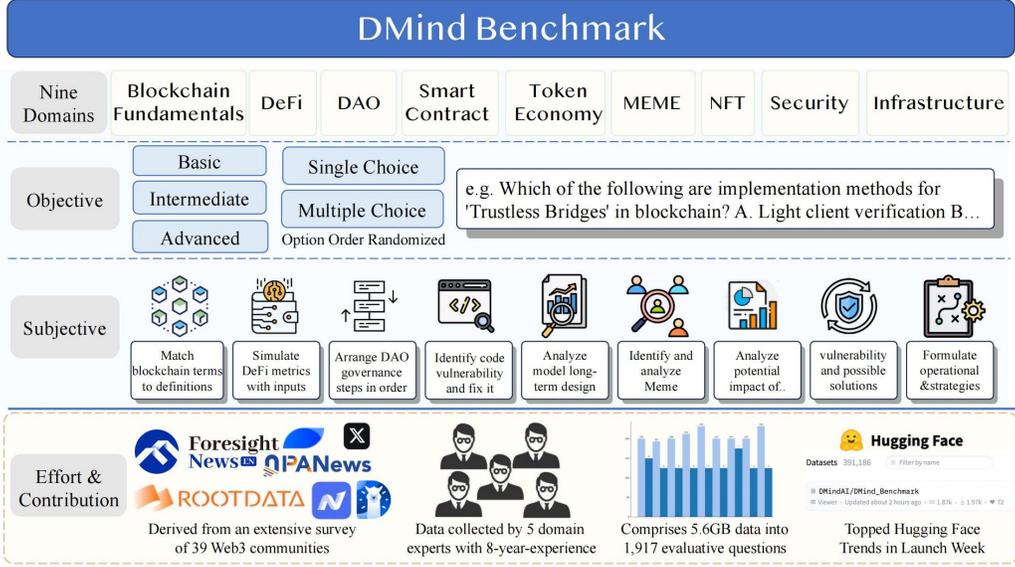
Figure 2: The DMind Benchmark framework, illustrating its nine evaluated Web3 domains, diverse objective and subjective task structures, and key metrics related to its development and community impact.

3.2 DMind Benchmark Assessment Design

Objective Assessment The objective assessment evaluates factual recall and basic understanding via multiple-choice questions across various web3 domains.

The score s_i for objective question Q_i is determined by its type $\tau(Q_i)$ (SC: Single-Choice, MC: Multiple-Choice), the model's selected options $A_M(Q_i)$, and correct options $C(Q_i)$ (where $|C(Q_i)| = 1$ for SC).

$$s_i = \mathbb{I}(\tau(Q_i) = \text{SC}) \cdot f_{\text{SC}}(A_M(Q_i), C(Q_i)) + \mathbb{I}(\tau(Q_i) = \text{MC}) \cdot f_{\text{MC}}(A_M(Q_i), C(Q_i)) \quad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function. Functions f_{SC} and f_{MC} are:

$$f_{\text{SC}}(A_M, C) = V_{\text{SC,corr}} \cdot \mathbb{I}(A_M = C \land |A_M| = 1) \quad (2)$$

$$f_{\text{MC}}(A_M, C) = V_{\text{MC,perf}} \cdot \mathbb{I}(A_M = C) + V_{\text{MC,part}} \cdot \mathbb{I}(\emptyset \neq A_M \subsetneq C) \quad (3)$$

Point values are $V_{\text{SC,corr}} = 2$ (correct SC), $V_{\text{MC,perf}} = 3$ (perfect MC), and $V_{\text{MC,part}} = 1$ (partial MC). Other outcomes yield 0 points. The formulae implement scoring rules: for SC, 2 pts for exact correct answer, 0 otherwise; for MC, 3 pts for perfect match, 1 for partial correctness (correct but incomplete selections), 0 if any incorrect option is chosen. Evaluation uses `test_objective.py`.

Subjective Assessment Subjective assessment gauges reasoning in complex web3 scenarios. It includes: (1) **Directly scored types** (e.g., Matching, Calculation) via output parsing; and (2) **AI-evaluated types** (e.g., Strategy Analysis, Code Audit) using Claude-3.7-sonnet for nuanced assessment.

For AI-evaluated types, the score s_j for question j uses a granular approach, formalized as:

$$s_j = \mathbf{w}_j^T \mathbf{e}_j = \sum_{k=1}^{p_j} w_{jk} \cdot e_{jk} \quad (4)$$

Here, $\mathbf{w}_j = (w_{jk})_{k=1}^{p_j}$ is the vector of predefined maximum points for p_j scoring elements in question j. $\mathbf{e}_j = (e_{jk})_{k=1}^{p_j}$ is the vector of corresponding normalized scores ($e_{jk} \in [0,1]$), with $e_{jk} = \text{Eval}_{\text{AI}}(A_{jk}, \mathcal{C}_{jk})$ based on the model's answer component A_{jk} and criteria $\mathcal{C}_{jk}$.

E.g., a 10-point question may have elements weighted 3, 3, 4. Claude evaluates each independently, ensuring comprehensive, weighted assessment. A keyword matching backup activates if AI evaluation fails. All types are handled by `test_subjective.py`.

By combining the scores from objective and subjective assessments, we can determine the final comprehensive score. The final score S_{total} combines objective ($S_{\text{obj}} = \sum s_i$) and subjective ($S_{\text{subj}} = \sum s_j$) scores. $S_{\text{obj,max}} = \sum s_{i,\text{max}}$ and $S_{\text{subj,max}} = \sum s_{j,\text{max}}$ are the respective maximums (where $s_{i,\text{max}} \in \{V_{\text{SC,corr}}, V_{\text{MC,perf}}\}$ and $s_{j,\text{max}} = \sum_k w_{jk}$).

The total score, S_{total}, is computed as:

$$S_{\text{total}} = \left(\omega_{\text{obj}} \cdot \tilde{S}_{\text{obj}} + \omega_{\text{subj}} \cdot \tilde{S}_{\text{subj}} \right) \cdot \mathcal{K}_{\text{scale}} \tag{5}$$

where:

- Normalized scores: $\tilde{S}_{\text{obj}} = S_{\text{obj}}/S_{\text{obj,max}}$, $\tilde{S}_{\text{subj}} = S_{\text{subj}}/S_{\text{subj,max}}$.
- Weights $\omega_{\text{obj}}, \omega_{\text{subj}}$ are proportions of sectional maximums to total maximum:

$$\omega_{\text{obj}} = \frac{S_{\text{obj,max}}}{S_{\text{obj,max}} + S_{\text{subj,max}}} \tag{6}$$

$$\omega_{\text{subj}} = \frac{S_{\text{subj,max}}}{S_{\text{obj,max}} + S_{\text{subj,max}}} \tag{7}$$

($\omega_{\text{obj}} + \omega_{\text{subj}} = 1$).
- $\mathcal{K}_{\text{scale}} = \frac{100}{9}$ is the scaling constant.

4 Experiments

To empirically assess the capabilities of contemporary Large Language Models (LLMs) within the Web3 domain and to demonstrate the utility of our **DMind Benchmark**, we performed a comprehensive evaluation. This section outlines our experimental setup, presents an overview of the general performance landscape including an overall model ranking, delves into a summarized analysis of model performance across specific Web3 subdomains, and concludes with key findings and their implications.

4.1 Experimental Setup

Our evaluation is anchored by the **DMind Benchmark**, which is designed to meticulously assess LLM proficiency across nine pivotal Web3 subfields: *(1) fundamental blockchain concepts (Fund.)*, *(2) blockchain infrastructure (Infra.)*, *(3) smart contract (S.C. Anal.)*, *(4) DeFi mechanisms (DeFi)*, *(5) Decentralized Autonomous Organizations (DAOs)*, *(6) Non-Fungible Tokens (NFTs)*, *(7) token economics (Token)*, *(8) meme concepts (Meme)*, and *(9) security vulnerabilities (Security)*. We evaluated a diverse set of 26 LLMs, broadly categorized into state-of-the-art (SOTA) models (detailed in Table 1) and Mini models (detailed in Table 2). This selection encompasses prominent model families such as Claude [1], DeepSeek [14], Gemini [46], GPT [38], Grok [58], and Qwen [2]. Model performance is quantified by accuracy scores (in percentages) for each subfield, facilitating a granular comparison. The color indicators in Table 1 and Table 2 provide a visual guide to performance tiers, as detailed in their respective captions.

To ensure reproducibility and a controlled evaluation environment, we standardized the generation parameters for querying the LLMs across the various tasks. For tasks requiring deterministic or factual outputs, such as multiple-choice questions and code-related analyses, a zero-shot prompting strategy was predominantly employed. Unless specific model architectures or particular task requirements dictated otherwise, we utilized the following decoding settings: a temperature of 0.75, a top-p (nucleus sampling) value of 0.9, and a top-k value of 20. The maximum number of new tokens to generate (max_tokens) was set to 16384, ensuring that responses were sufficiently comprehensive without being overly verbose. These parameters were selected to foster coherent and accurate responses while minimizing undesired output randomness, thereby allowing for a more direct comparison of the models' inherent capabilities on the benchmark tasks.

To ensure the robustness of our results, all models were evaluated five times. The median score from these five runs was taken as the final reported performance for each subfield. Error margins, depicted by error bars in the accompanying bar charts, represent the score variability across these runs and consistently remained within $\pm 1.5\%$ for all models.
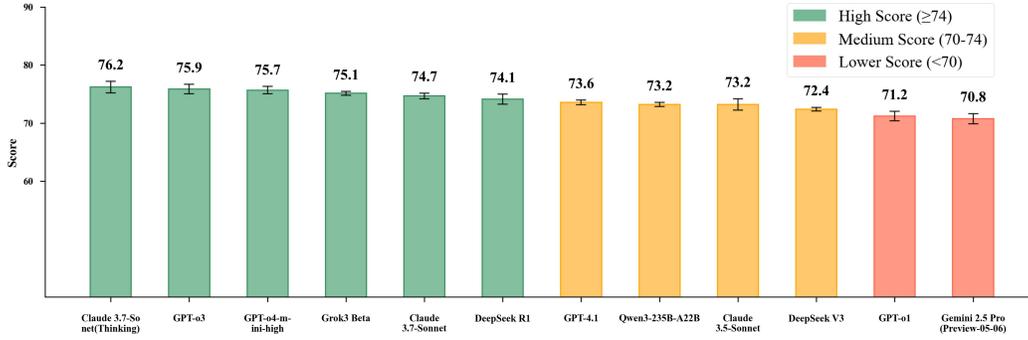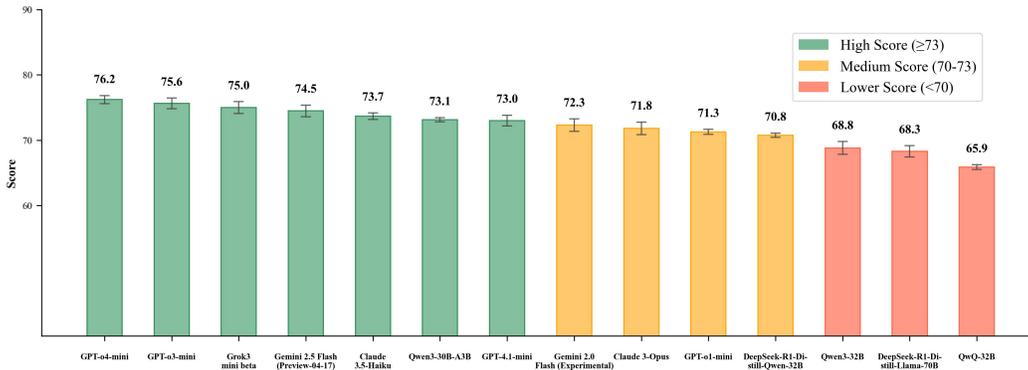
Figure 3: Scores of Sota Models

Figure 4: Scores of Mini Models

4.2 Overall Performance Ranking and General Trends

To provide a holistic view of model capabilities, Figure 3 and Figure 4 presents the overall performance ranking of all 26 evaluated LLMs, based on their aggregate scores across the nine subdomains of the DMind Benchmark. This ranking reveals distinct performance tiers among the models. Leading SOTA models, such as Claude 3.7-Sonnet (Thinking), GPT-o3, and GPT-o4-mini-high, consistently occupy the upper echelon, showcasing broad and robust understanding across multiple facets of Web3. Notably, several Mini models, including GPT-o4-mini and Grok3 mini beta, also feature prominently in the higher rankings, demonstrating that optimized smaller models can be highly competitive.

The overall scores detailed in Table 1 and Table 2, confirm general trends: a widespread proficiency in foundational areas like *Blockchain Infrastructure* and *Fundamental Blockchain Concepts* across most models. Conversely, the consistently low scores in highly specialized fields such as *Token Economics* significantly impacted the overall standings for many, underscoring this as a critical area for improvement. The distribution of overall scores further highlights the varying degrees of specialization and generalization among current LLMs when applied to the complex Web3 domain.

4.3 Performance Insights Across Web3 Subdomains

The DMind Benchmark allows for a granular analysis of LLM capabilities within each specific Web3 subdomain. Below, we provide a high-level summary of performance trends, with detailed scores available in Table 1 and Table 2.

1. Fundamental Blockchain Concepts (Fund.): This area was a clear strength for most LLMs. Both SOTA and leading Mini models consistently achieved high accuracy, often exceeding 90%, indicating a solid grasp of core blockchain principles.

Table 1: Performance scores (%), normalized to a total of 100, of State-of-the-Art (SOTA) LLMs on the DMind Benchmark across nine Web3 subdimensions (Fund., Infra., S.C. Anal., DeFi, DAOs, NFTs, Token, Meme, Sec.). Cell background colors denote performance levels: Green for scores ≥ 90%, Yellow for scores in the range [80%, 90%), and Red for scores < 80%.

Model	Fund.	Infra.	S.C.	DeFi	DAOs	NFTs	Token	Meme	Sec.
Claude 3.7-Sonnet (Thinking)	88.66	97.96	91.37	83.05	73.32	85.17	25.40	68.33	72.66
Claude 3.7-Sonnet	89.69	94.97	89.67	83.06	73.32	81.80	24.80	63.70	71.18
Claude 3.5-Sonnet	89.28	94.85	87.50	80.85	71.69	80.45	24.40	62.50	67.36
DeepSeek R1	91.55	97.03	82.83	82.65	72.78	79.64	22.80	69.44	68.40
DeepSeek V3	90.31	95.81	83.00	77.55	73.68	74.35	23.80	63.70	69.44
Gemini 2.5 Pro (Preview-05-06)	81.03	93.66	81.37	78.16	67.88	76.87	19.40	67.96	70.49
GPT-o4-mini-high	91.75	98.57	89.02	83.26	74.05	81.07	23.00	74.63	64.80
GPT-o3	92.99	98.36	88.43	81.02	74.59	80.52	24.20	71.67	71.01
GPT-o1	90.31	98.36	89.31	83.06	68.24	69.71	23.40	51.11	67.45
GPT-4.1	88.87	97.55	87.45	77.35	73.14	75.60	22.40	70.19	69.62
Grok3 beta	90.72	96.52	88.08	81.26	69.87	80.69	24.00	73.70	71.35
Qwen3-235B A22B	88.66	95.60	79.88	79.39	75.32	79.73	21.40	70.56	68.40

Table 2: Performance scores (%), normalized to a total of 100, of Mini LLMs on the DMind Benchmark across nine Web3 subdimensions (Fund., Infra., S.C. Anal., DeFi, DAOs, NFTs, Token, Meme, Sec.). Cell background colors denote performance levels: Green for scores ≥ 85%, Yellow for scores in the range [70%, 85%), and Red for scores < 70%.

Model	Fund.	Infra.	S.C.	DeFi	DAOs	NFTs	Token	Meme	Sec.
Claude 3.5-Haiku	91.13	96.32	86.08	75.46	72.05	83.22	24.40	63.89	70.57
Claude 3-Opus	83.51	91.72	78.82	77.55	72.23	77.73	24.60	69.44	70.75
DeepSeek-R1-Distill-Llama-70B	83.71	95.40	82.35	80.81	66.06	65.96	24.20	49.44	66.75
DeepSeek-R1-Distill-Qwen-32B	83.51	92.43	77.25	76.32	72.05	75.61	22.40	70.37	67.10
Gemini 2.5 Flash (Preview-04-17)	88.45	97.03	82.94	80.20	73.50	82.52	22.80	71.67	71.35
Gemini 2.0 Flash (Experimental)	85.15	94.89	81.37	79.57	71.51	77.65	21.80	68.89	69.01
GPT-o4-mini	91.34	97.96	88.23	82.85	74.05	78.60	25.20	73.52	73.61
GPT-o3-mini	91.96	98.16	86.08	81.63	71.14	84.18	23.60	69.44	74.48
GPT-o1-mini	87.63	95.50	80.35	76.32	69.51	74.92	23.40	64.63	69.18
GPT-4o-mini	82.06	86.50	75.88	76.68	68.06	73.66	22.40	60.74	67.19
Grok3 mini beta	89.69	97.75	84.02	83.47	74.05	79.99	23.40	69.07	73.44
Qwen3-32B	89.69	97.96	78.05	79.50	66.97	70.70	25.20	49.63	61.63
Qwen3-30B-A3B	88.45	96.93	78.63	80.20	74.23	78.55	23.20	69.81	68.23
QwQ-32B	82.69	91.21	73.35	73.06	67.88	69.38	22.20	47.04	66.15

2. Blockchain Infrastructure (Infra.): Models demonstrated the highest proficiency in this subdomain. Nearly all evaluated LLMs, irrespective of size, exhibited excellent understanding, with many SOTA and Mini models scoring above 95%.

3. smart contract (S.C. Anal.): Performance was more varied here. Top-tier SOTA and several standout Mini models showed strong capabilities (scores often in the 80 − 90% range). However, this remains a challenging area for a broader set of models, reflecting the complexity of code understanding and debugging.

4. DeFi Mechanisms (DeFi): Leading models generally performed well, achieving scores typically around 80 − 85%. This suggests a good understanding of core DeFi concepts by the more capable LLMs, though the rapid evolution and intricacies of DeFi continue to pose challenges.

5. Decentralized Autonomous Organizations (DAOs): This subfield proved moderately difficult for most LLMs. Scores typically clustered in the 70 − 75% range, indicating a partial but not comprehensive understanding of DAO governance and operational dynamics.

6. Non-Fungible Tokens (NFTs): Proficiency in NFTs was mixed. While some SOTA and Mini models achieved scores above 80%, others showed more limited understanding, suggesting variability in grasping NFT standards, use cases, and cultural contexts.

7. Token Economics (Token): This was unequivocally the most challenging subdomain for all LLMs. Scores were uniformly low across both SOTA and Mini models, rarely exceeding 26%. This highlights a critical gap in current LLMs' ability to reason about complex cryptoeconomic systems.

8. Meme Concepts (Meme): Understanding Web3-specific meme concepts presented considerable difficulty. While a few models approached scores in the mid-70s, most struggled, with performance often in the $50 - 65\%$ range, reflecting the challenge of capturing nuanced cultural and rapidly evolving information.

9. Security Vulnerabilities (Sec.): This critical area showed moderate performance overall. Scores for most models, including SOTA, were typically in the $60 - 75\%$ range. While some leading Mini models performed competitively, the general level of proficiency indicates a significant need for improvement in identifying and reasoning about security flaws.

4.4 Key Findings and Implications

Our comprehensive evaluation using the DMind Benchmark, encompassing an overall performance ranking (Figure 3 and Figure 4) and detailed subdomain analysis (Table 1 and Table 2), yields several crucial insights:

Stratified Performance and Heterogeneous Proficiency: The Web3 LLM landscape reveals distinct performance tiers. While large SOTA models generally lead, highly optimized Mini models are notably competitive, often excelling in specific subdomains and underscoring their potential. This varied proficiency shows strong capabilities in foundational areas like infrastructure and basic concepts, contrasting with diminished performance in more specialized, dynamic, or interdisciplinary subdomains (e.g., DAOs, NFTs, meme concepts) which demand more sophisticated understanding.

Persistent Challenges and Subdomain-Specific Difficulties: Persistent challenges highlight critical areas for LLM improvement. *Token Economics* stands out as a profound and universal difficulty. Accurately navigating *Security Vulnerabilities* and the nuances of *smart contract* remain significant hurdles, alongside understanding other complex areas like DAOs, NFTs, and meme concepts, requiring targeted research and development.

DMind Benchmark as a Catalyst for Progress: These findings validate the DMind Benchmark as an essential instrument for the AI and Web3 communities. It provides a standardized, granular framework for assessing LLM proficiency, identifying critical areas for improvement, and ultimately guiding the development of more robust, reliable, and specialized AI solutions tailored for the complexities of the decentralized web.

In conclusion, while the integration of LLMs into Web3 is rapidly advancing, achieving true expertise and reliability across its diverse and evolving landscape necessitates focused innovation. The DMind Benchmark offers a clear path for measuring progress and steering the development of next-generation LLMs equipped to tackle these challenges.

5 Conclusion

This paper introduced the **DMind Benchmark**, a comprehensive evaluation suite designed to address the critical need for assessing Large Language Model (LLM) proficiency in the complex Web3 domain. Our evaluation of 26 LLMs across nine Web3 subfields revealed strong performance in foundational areas but significant weaknesses in specialized topics like token economics and security vulnerabilities, highlighting specific avenues for LLM development.

The DMind Benchmark serves as a valuable instrument for probing LLM capabilities beyond general natural language understanding, offering insights into how these models handle the interdisciplinary and rapidly evolving nature of Web3. By systematically pinpointing current limitations, particularly in areas requiring deep domain-specific reasoning, our work helps inform efforts to enhance LLM robustness and applicability in specialized, high-stakes contexts. The benchmark's open-source availability and positive community reception suggest its potential to foster collaborative research and steer the development of more adept AI systems for such challenging domains.

While DMind Benchmark offers a significant step, its continued relevance will depend on ongoing updates to reflect Web3's dynamism. Future work may involve expanding task diversity and exploring multimodal evaluations. We believe the DMind Benchmark will not only aid in developing more capable LLMs for Web3 but also offer a model for creating nuanced evaluation frameworks in other specialized AI application areas, thereby fostering the continuous improvement of LLM capabilities in specialized domains.

References

[1] Anthropic. Claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, 2024.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, and et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] M. Bartoletti and L. Pompianu. An empirical analysis of smart contracts: platforms, applications, and design patterns. In *Financial Cryptography and Data Security: FC 2017 International Workshops*, Lecture Notes in Computer Science, pages 494–509. Springer International Publishing, 2017.

[4] R. Belchior, A. Vasconcelos, S. Guerreiro, and M. Correia. A survey on blockchain interoperability: Past, present, and future trends. *ACM Computing Surveys*, 54(8):1–41, 2021.

[5] Cristiano Bellavitis, Christian Fisch, Paul P. Momtaz, and et al. The rise of decentralized autonomous organizations (daos): a first empirical glimpse. *Venture Capital*, 2023.

[6] V. Buterin. Ethereum: A next-generation smart contract and decentralized application platform. Ethereum White Paper, 2014.

[7] Vitalik Buterin et al. Ethereum white paper. *GitHub repository*, 1(22-23):5–7, 2013.

[8] Christian Catalini, Alonso de Gortari, and Nihar Shah. Some simple economics of stablecoins. *Annual Review of Financial Economics*, 14(1):117–135, 2022.

[9] Siddhartha Chatterjee and Bina Ramamurthy. Efficacy of various large language models in generating smart contracts. In Kohei Arai, editor, *Advances in Information and Communication*, pages 482–500, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-85363-0.

[10] Mark Chen, Jerry Tworek, Jun, and et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[11] Qianqian Chen, Wen Han, Zhihao Chen, and et al. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743, 2024.

[12] Yan Chen and Cristiano Bellavitis. Blockchain disruption and decentralized finance: The rise of decentralized business models. *Journal of Business Venturing Insights*, 13:e00151, 2020.

[13] K. R. Dearstyne, A. D. Rodriguez, and J. Cleland-Huang. Supporting software maintenance with dynamically generated document hierarchies. In *ICSME*, pages 426–437, 2024.

[14] DeepSeek-AI, Daya Guo, Dejian Yang, and et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[15] Caleb Geren, Amanda Board, Gaby G. Dagher, Tim Andersen, and Jun Zhuang. Blockchain for large language model security and safety: A holistic survey. *SIGKDD Explor. Newsl.*, 26(2): 1–20, January 2025. ISSN 1931-0145. doi: 10.1145/3715073.3715075.

[16] Neel Guha, Julian Nyarko, Daniel Ho, and et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc., 2023.

[17] Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, and et al. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*, 2024.

[18] Samer Hassan and Primavera De Filippi. Decentralized autonomous organization. *Internet Policy Review*, 10(2), 2021.

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[20] Md Rafiqul Islam, Muhammad Mahbubur Rahman, Md Mahmud, Mohammed Ataur Rahman, Muslim Har Sani Mohamad, and Abd Halim Embong. A review on blockchain security issues and challenges. In *2021 IEEE 12th control and system graduate research colloquium (ICSGRC)*, pages 227–232. IEEE, 2021.

[21] Kensuke Ito. Cryptoeconomics and tokenomics as economics: A survey with opinions. *arXiv preprint arXiv:2407.15715*, 2024.

[22] Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, and et al. Big-bench extra hard. *arXiv preprint arXiv:2502.19187*, 2025.

[23] Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. Medexqa: Medical question answering benchmark with multiple explanations. *arXiv preprint arXiv:2406.06331*, 2024.

[24] David Krause. Beyond the hype: A meme coin reality check for retail investors. *Available at SSRN 4891841*, 2024.

[25] Ennan Lai and Wenjun Luo. Static analysis of integer overflow of smart contracts in ethereum. In *ICCSP*, 2020.

[26] Y. Li, B. Luo, and Q. et al. Wang. A reflective llm-based agent to guide zero-shot cryptocurrency trading. *arXiv preprint arXiv:2407.09546*, 2024.

[27] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, and et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2023.

[28] Chao Liu, Han Liu, Zhao Cao, Zhong Chen, Bangdao Chen, and Bill Roscoe. Reguard: finding reentrancy bugs in smart contracts. In *ICSE*, 2018.

[29] Zefang Liu1, Jialei Shi1, and John F. Buford1. Cyberbench: A multi-task benchmark for evaluating large language models in cybersecurity. In *AICS*. Curran Associates, Inc., 2024.

[30] Hou-Wan Long, Hongyang Li, and Wei Cai. Coinclip: A multimodal framework for evaluating the viability of memecoins in the web3 ecosystem. *arXiv preprint arXiv:2412.07591*, 2024.

[31] M.S. Looijenga. Rechtbert : Training a dutch legal bert model to enhance legaltech, December 2024. URL http://essay.utwente.nl/104811/.

[32] H. Luo, J. Luo, and A. V. Vasilakos. Bc4llm: A perspective of trusted artificial intelligence when blockchain meets large language models. *Neurocomputing*, 599:128089, 2024.

[33] V. Mothukuri, R. M. Parizi, and J. L. et al. Massa. An ai multi-model approach to defi project trust scoring and security. In *Blockchain*, pages 19–28, 2024.

[34] Matthieu Nadini, Laura Alessandretti, Flavio Di Giacinto, Mauro Martino, Luca Maria Aiello, and Andrea Baronchelli. Mapping the nft revolution: market trends, trade networks, and visual features. *Scientific reports*, 11(1):20902, 2021.

[35] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.

[36] D. Nam, A. Macvean, and V. et al. Hellendoorn. Using an llm to help with code understanding. In *ICSE*, pages 1–13, 2024.

[37] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

[38] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, and Sam Altman et al. Gpt-4 technical report, 2024.

[39] Peterson K. Ozili. Decentralized finance research and developments around the world. *Journal of Banking and Financial Technology*, 2022.

[40] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*. PMLR, 2022.

[41] Ruhul Saha, Gaurav Kumar, Mauro Conti, and Sujata Pal. Dhacs: Smart contract-based decentralized hybrid access control for industrial internet-of-things. *IEEE Transactions on Industrial Informatics*, 18(5):3452–3461, 2021.

[42] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, and et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[43] Karan Singhal, Shekoofeh Azizi, Tu, and et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 8 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2.

[44] M. Suri, P. Mathur, and F. et al. Dernoncourt. Docedit-v2: Document structure editing via multimodal llm grounding. In *EMNLP*, pages 15485–15505, 2024.

[45] N. Szabo. Formalizing and securing relationships on public networks. *First Monday*, 2(9), 1997.

[46] G. Team, R. Anil, S. Borgeaud, and et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[47] Gemma Team, Thomas Mesnard, Cassidy Hardin, and et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, and et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[49] K. Toyoda, X. Wang, and M. et al. Li. Blockchain data analysis in the era of large-language models. *arXiv preprint arXiv:2412.09640*, 2024.

[50] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[51] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2020.

[52] Qin Wang, Rujia Li, Qi Wang, and Shiping Chen. Non-fungible token (nft): Overview, evaluation, opportunities and challenges. *arXiv preprint arXiv:2105.07447*, 2021.

[53] Shuai Wang, Wenwen Ding, Juanjuan Li, Yong Yuan, Liwei Ouyang, and Fei-Yue Wang. Decentralized autonomous organizations: Concept, model, and applications. *IEEE Transactions on Computational Social Systems*, 6(5):870–878, 2019.

[54] Sam Werner, Daniel Perez, Lewis Gudgeon, Ariah Klages-Mundt, Dominik Harz, and William Knottenbelt. Sok: Decentralized finance (defi). In *Proceedings of the 4th ACM Conference on Advances in Financial Technologies*, pages 30–46, 2022.

[55] Gavin Wood et al. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper*, 151(2014):1–32, 2014.

[56] C. Wu, J. Chen, and Z. et al. Wang. Semantic sleuth: Identifying ponzi contracts via large language models. In *ASE*, pages 582–593, 2024.

[57] Shijie Wu, Ozan Irsoy, Lu, and et al. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[58] xAI. About grok, 2025. URL `https://grok.online/zh/about`.

[59] Dylan Yaga, Peter Mell, Nik Roby, and Karen Scarfone. Blockchain technology overview. *arXiv preprint arXiv:1906.11078*, 2019.

[60] J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander. Where is current research on blockchain technology?—a systematic review. *PloS one*, 11(10):e0163477, 2016.

[61] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, and et al. Safetybench: Evaluating the safety of large language models. *arXiv preprint arXiv:2309.07045*, 2024.

[62] L. Zhong and Z. Wang. Can llm replace stack overflow? a study on robustness and reliability of large language model code generation. In *AAAI*, 2024.

[63] Q. Zhou, H. Huang, Z. Zheng, and J. Bian. Solutions to scalability of blockchain: A survey. *IEEE Access*, 8:16440–16455, 2020.

A Representative Evaluation Items for Each Category

To make the design philosophy of DMind Benchmark more transparent, we highlight one carefully–selected item from every category contained in `categorized_questions.md`. For each item we discuss (a) why it is emblematic of its Web3 sub–field, (b) which fine–grained capabilities it probes in large language models, and (c) the typical failure modes we observed during internal evaluation.

A.1 Blockchain Fundamentals — Consensus Mechanisms

> **Question**
>
> "Which consensus mechanism has received more attention in terms of energy efficiency?"
> Options: (A) Proof of Work (PoW) (B) Proof of Stake (PoS) (C) Proof of Capacity (PoC) (D) Proof of Importance (PoI).
> **Correct**: (B) *Knowledge point*: Blockchain Basics — Consensus Mechanisms.

> **Representativeness**
>
> This question targets the foundational understanding of consensus mechanisms' core properties, specifically focusing on energy efficiency—a critical motivation behind the Ethereum Merge and many subsequent blockchain designs.

> **LLM Competence Dimensions**
>
> - Factual recall of basic consensus properties and their comparative advantages
> - Ability to connect technical mechanisms with their real-world engineering implications
> - Discrimination between primary and secondary characteristics of blockchain systems

> **Model Performance**
>
> Both SOTA and Mini models demonstrated strong performance (>90% accuracy), indicating that the fundamental concepts of consensus mechanisms are well-represented in training data. However, we observed that Mini models occasionally confused PoS with PoC, suggesting some conflation of efficiency-focused consensus variants.

> **Task**
>
> "Explain the principles, advantages, disadvantages, and applicable scenarios of different types of consensus mechanisms (PoW, PoS, DPoS, etc.)"
> **Format**: Matching comparison requiring structured analysis across multiple dimensions.

Representativeness

This task requires comprehensive knowledge spanning technical principles, economic incentives, and practical implementation considerations—testing both breadth and depth of understanding about the cornerstone technologies of blockchain systems.

LLM Competence Dimensions

- Multidimensional comparative analysis across technical mechanisms
- Causal reasoning about trade-offs between security, decentralization, and efficiency
- Synthesis of technical attributes with practical deployment considerations

Model Performance

SOTA models like Claude 3.7-Sonnet and GPT-o3 provided thorough analyses covering 80% or more of the scoring criteria, with clear discussion of energy consumption, decentralization impacts, and appropriate use cases. Mini models typically focused on singular dimensions (often energy efficiency alone) while neglecting deeper security trade-offs, demonstrating the gap in complex analytical capabilities between model tiers.

A.2 Blockchain Infrastructure — Scaling Solutions

Question

"What main problem does blockchain sharding technology solve?"
Options: (A) Security (B) Scalability (C) Decentralization (D) Anonymity.
Correct: (B) *Knowledge point*: Layer1 — Off-chain Scaling — Sharding Technology.

Representativeness

Sharding is the canonical answer to the "single-chain bottleneck" and sits at the heart of the security-decentralization-scalability trilemma that defines modern Layer-1 R&D.

LLM Competence Dimensions

- Technical categorization of blockchain scaling approaches
- Understanding of the primary value proposition of horizontal scaling
- Recognition of core blockchain design constraints and their relationships

Model Performance

This question saw exceptionally high correct response rates (approximately 97%), making it the best-performing category overall. This suggests that infrastructure-related concepts, particularly regarding scalability solutions, are well-represented in the models' training data.

Task

"Analyze and compare Layer 1 blockchains like Solana (PoH consensus), Aptos/Sui (Move language), and Monad (enhanced EVM compatibility). Evaluate their technical mechanisms, ecosystem development, and formulate investment strategies."
Format: Strategy analysis requiring technical evaluation and practical recommendations.

This task requires synthesizing knowledge across technical architecture, token economics, and ecosystem dynamics—a real-world scenario that tests both technical depth and strategic breadth.

LLM Competence Dimensions

- Cross-protocol comparative analysis of diverse technical stacks
- Evaluation of tokenomics models and their market implications
- Strategic reasoning connecting technical capabilities to investment thesis

Model Performance

Only top-tier models like GPT-o3 and Claude 3.7-Sonnet provided complete analyses across all three required dimensions (technical mechanisms, ecosystem assessment, and investment strategy). Most Mini models demonstrated adequate technical analysis but showed notable weaknesses in tokenomics evaluation, often providing generic comments about "token utility" without specific mechanisms or distribution metrics. This illustrates that vertical domain integration (connecting technology to economics) remains a distinguishing capability of larger models.

A.3 Smart Contracts — Security and Optimization

> **Question**
>
> "Which variable type in Solidity is used to store Ether amounts?"
> Options: (A) uint (B) int (C) wei (D) ether.
> **Correct**: (A) *Knowledge point*: Solidity Fundamentals — Data Types.

Representativeness

This question probes fundamental knowledge of Solidity's type system and the representation of native assets—essential knowledge for anyone working with smart contracts.

LLM Competence Dimensions

- Recall of programming language specifics in blockchain contexts
- Understanding of the relationship between native tokens and their programmatic representation
- Distinction between units of measurement and data types

Model Performance

Most models performed well on this question (>85% accuracy), demonstrating solid grasp of basic Solidity syntax. However, performance declined sharply on more complex questions about delegate calls, ABI encoding, and gas optimization, validating our progressive difficulty design in the benchmark.

> **Task**
>
> "Identify the vulnerability in the smart contract and provide fixed code for VulnerableBank.sol containing a reentrancy vulnerability."
> **Format**: Code audit requiring identification, explanation, and correction of contract flaws.

Representativeness

Reentrancy attacks represent one of the most notorious and costly vulnerabilities in smart contract history. This task simulates a critical security audit scenario requiring both identification and remediation.

LLM Competence Dimensions

- Code comprehension and vulnerability detection
- Knowledge of secure coding patterns (Checks-Effects-Interactions)
- Implementation of appropriate security controls (ReentrancyGuard)
- Ability to generate syntactically valid and functionally correct code

Model Performance

Models with strong code generation capabilities like GPT-o4-mini-high and GPT-o3 could identify the vulnerability, explain the attack vector, and implement complete fixes that passed automated testing. In contrast, models like DeepSeek R1 typically described the vulnerability correctly but provided incomplete or non-compiling code fixes, highlighting the gap between conceptual understanding and practical implementation capabilities.

A.4 DeFi Mechanisms — Financial Models and Calculations

> **Question**
> "What does AMM stand for?"
> Options: (A) Automated Market Maker (B) Advanced Market Management (C) Automated Money Market (D) Asset Management Model.
> **Correct**: (A) *Knowledge point*: DeFi Basics — Market Structure.

Representativeness

AMMs represent a fundamental innovation in DeFi, replacing traditional order books with algorithmic liquidity provision—a cornerstone concept for understanding decentralized exchanges.

LLM Competence Dimensions

- Recognition of domain-specific terminology and acronyms
- Understanding of core DeFi infrastructure components
- Differentiation between similar financial concepts

Model Performance

Approximately 83% of models answered correctly, with errors concentrated in smaller models, suggesting that accurate recall of domain-specific terminology correlates with parameter count even for seemingly simple acronym expansions.

> **Task**
>
> "Calculate the Ethereum price at liquidation given a BTC collateral value of $50M with collateral ratio 0.8, liquidation threshold 0.83, initial BTC price $85,000, initial ETH price $2,200, and BTC price at liquidation $84,000."
> **Format**: Numerical reasoning requiring multiple calculation steps.

Representativeness

This task mimics real-world DeFi risk assessment scenarios where understanding collateralization ratios, liquidation thresholds, and price relationships is critical for predicting market events.

LLM Competence Dimensions

- Multi-step mathematical reasoning
- Application of financial formulas in cryptocurrency contexts
- Precise calculation with appropriate rounding and units

Model Performance

Claude 3.7 and GPT-o4 variants consistently calculated the exact answer ($2,255.56) with appropriate decimal formatting. Most Mini models made calculation errors, particularly by missing the liquidation threshold coefficient (0.83), resulting in >5% error. This demonstrates that complex numerical reasoning with multiple variables remains challenging for smaller models, and highlights the benchmark's effectiveness at differentiating mathematical capabilities.

A.5   Decentralized Autonomous Organizations — Governance Models

> **Question**
>
> "What is the main purpose of governance tokens in DAOs?"
> Options: (A) Paying transaction fees (B) Participating in protocol decisions (C) Acting as stablecoins (D) Cross-chain transactions.
> **Correct**: (B) *Knowledge point*: DAO Basics — Governance Mechanisms.

Representativeness

Governance tokens embody the core mechanism through which DAOs enable decentralized decision-making—a defining feature that distinguishes them from traditional organizations.

LLM Competence Dimensions

- Understanding of token utility frameworks
- Recognition of governance as a primary token function
- Differentiation between various token use cases

Model Performance

Overall accuracy in the DAO category averaged 72%, substantially higher than the Token Economics category (24%), revealing a notable disparity in understanding between governance concepts and economic mechanisms despite their related nature.

Task

"Explore blockchain governance models and their impact on decentralization. Compare different approaches (on-chain vs. off-chain, formal vs. informal) and their advantages and disadvantages."
Format: Comparative analysis requiring evaluation of governance trade-offs.

Representativeness

This task examines the theoretical and practical dimensions of blockchain governance, targeting the central tension between decentralization ideals and operational efficiency.

LLM Competence Dimensions

- Analysis of governance structures and their implications
- Evaluation of power distribution in token-weighted systems
- Recognition of centralization risks in different governance approaches

Model Performance

Top-scoring responses required comprehensive analysis of both token-weighted voting systems and the role of core developers in governance. Mini models typically addressed only the most visible aspects (on-chain voting) while neglecting the "formal vs. informal" governance dimension, scoring approximately 50% on average. This indicates that nuanced understanding of governance structures remains challenging for smaller models.

A.6 Non-Fungible Tokens — Standards and Applications

Question

"Which project first popularized the ERC-721 NFT standard?"
Options: (A) CryptoKitties (B) CryptoPunks (C) OpenSea (D) Bored Ape Yacht Club.
Correct: (A) *Knowledge point*: NFT History — Standard Evolution.

Representativeness

Understanding the historical development of NFT standards provides insight into how technological innovations emerge and evolve in the Web3 space.

LLM Competence Dimensions

- Recall of blockchain technology history and milestones
- Distinction between standards, implementations, and platforms
- Chronological ordering of Web3 developments

Model Performance

SOTA models averaged approximately 80% accuracy on NFT questions, while Mini models achieved around 70%. Common errors included selecting OpenSea, indicating confusion between standard creation and marketplace adoption. This suggests limitations in models' understanding of the relationship between protocols and applications built on top of them.

Task

"Evaluate the effectiveness and applications of blockchain privacy protection technologies for NFTs. Compare zero-knowledge proofs, coin mixing, and ring signatures, analyzing their level of privacy, computational overhead, and appropriate use cases."
Format: Technical comparison requiring security and privacy analysis.

Representativeness

This task explores the intersection of NFTs with privacy technologies, a critical consideration for digital asset applications in contexts requiring confidentiality or regulatory compliance.

LLM Competence Dimensions

- Comparative analysis of cryptographic privacy techniques
- Understanding of privacy-transparency trade-offs
- Evaluation of computational efficiency and implementation complexity

Model Performance

Only Claude 3.7-Sonnet consistently achieved scores above 8/10, providing balanced analysis of both technical mechanisms and their practical implications. Most models omitted discussion of regulatory compliance considerations, resulting in scoring penalties. This highlights the challenge of integrating technical, legal, and practical dimensions in complex domain analyses.

A.7 Token Economics — Monetary Design and Incentives

Question

"What does 'Collateralization Ratio' refer to in DeFi?"
Options: (A) The ratio of loan amount to collateral value (B) Protocol yield rate (C) Transaction fee rate (D) Liquidity ratio.
Correct: (A) *Knowledge point*: Token Economics — Financial Ratios.

Representativeness

Collateralization ratios form the foundation of risk management in decentralized lending, a critical concept bridging traditional finance and crypto-native mechanisms.

LLM Competence Dimensions

- Understanding of financial metrics in DeFi contexts
- Knowledge of risk management principles in collateralized lending
- Ability to identify the correct mathematical relationship

Model Performance

This category showed the lowest overall performance (<25% accuracy), with many models reversing the ratio direction. This striking underperformance across all model tiers suggests a systematic gap in training data or conceptual understanding of DeFi mathematical concepts, highlighting a critical area for improvement.

> **Task**
>
> "Analyze a newly launched DeFi platform token's economic model and long-term sustainability. Consider token distribution, utility, inflation/deflation mechanisms, and governance rights."
> **Format**: Fill-in-the-blank assessment requiring comprehensive tokenomics analysis.

Representativeness

This task evaluates understanding of token economic design principles that determine long-term value accrual and distribution—essential knowledge for evaluating Web3 projects.

LLM Competence Dimensions

- Analysis of token distribution models and vesting schedules
- Evaluation of utility mechanisms and value capture
- Understanding of inflationary/deflationary dynamics

Model Performance

Most models provided generic discussions of "value accrual" without specific mechanisms, scoring below 40% across the inflation/deflation, governance, and utility dimensions. Even SOTA models struggled to articulate concrete tokenomics principles beyond surface-level observations, confirming token economics as a significant knowledge gap in current LLMs.

A.8 Meme Concepts — Cultural Narratives and Community

> **Question**
>
> "Which slogan is associated with Dogecoin?"
> Options: (A) "Much Wow" (B) "WAGMI" (C) "To the Moon" (D) "Wen Lambo".
> **Correct**: (A) *Knowledge point*: Meme Culture — Community Slogans.

Representativeness

Meme tokens represent a fusion of internet culture with tokenized value, making cultural literacy essential for understanding community-driven projects.

LLM Competence Dimensions

- Recognition of cultural references in crypto communities
- Association of specific phrases with particular projects
- Understanding of internet culture's influence on token communities

Model Performance

High-parameter models achieved >70% accuracy, leveraging their exposure to social media content, while Mini models fell to approximately 50%, indicating that cultural knowledge correlates strongly with training data volume and recency.

Task

"Analyze GameFi operational strategies by comparing Pixelmon (NFT-first approach), Treasure DAO (ecosystem around MAGIC token), and Apeiron (three-token model). Evaluate their economic models, player engagement approaches, and provide recommendations."
Format: Case study analysis requiring marketing and economic assessment.

Representativeness

This task examines the intersection of gaming, community building, and token economics—a prime example of how Web3 projects leverage cultural engagement for economic activity.

LLM Competence Dimensions

- Comparative analysis of community-driven projects
- Understanding of gaming economy design principles
- Integration of cultural narratives with economic incentives

Model Performance

Only GPT-o3 provided a comprehensive analysis that systematically captured the trend toward "play-first, earn-later" models and discussed specific mechanisms like multi-token designs to control inflation. Most models offered generic advice without connecting cultural elements to economic sustainability, revealing limitations in cross-domain integration.

A.9 Security Vulnerabilities — Risk Assessment and Mitigation

Question

"Flash loan exploits primarily target which vulnerability?"
Options: (A) Price oracle manipulation (B) Liquidity shortages (C) Governance attacks (D) Sybil attacks.
Correct: (A) *Knowledge point*: Security — Attack Vectors and Exploits.

Representativeness

Flash loan attacks represent some of the most devastating exploits in DeFi history, making understanding their mechanics crucial for security assessment.

LLM Competence Dimensions

- Identification of attack patterns and vulnerability classes
- Understanding of temporal dependencies in smart contract execution
- Knowledge of oracle design and manipulation vectors

Model Performance

Average accuracy was approximately 65%, reflecting incomplete understanding of specific attack chains. Models often confused the primary vector (price oracle manipulation) with secondary aspects like governance or liquidity issues, indicating difficulties in distilling primary causal relationships in complex attack scenarios.

Representativeness

This task simulates real-world security auditing, requiring both technical understanding of vulnerabilities and practical judgment about their severity and remediation.

LLM Competence Dimensions

- Classification of security vulnerabilities by type and impact
- Risk assessment and prioritization of mitigation efforts
- Technical recommendations for vulnerability remediation

Model Performance

Claude 3.7-Sonnet demonstrated strong capabilities, accurately classifying vulnerabilities, assigning appropriate severity levels, and providing specific mitigation strategies. Most models could enumerate vulnerabilities but struggled with proper classification and prioritization, averaging around 4/10 points. This suggests that structured security assessment remains challenging for most LLMs, requiring further refinement in threat modeling capabilities.

B   Resource Consumption Guidelines

For our comprehensive evaluation, we selected 12 representative LLMs and meticulously tracked token consumption metrics across all models. As shown in Table B, token utilization varies significantly between models, with Gemini-2.5 Pro (Preview-05-06) consuming the highest total tokens (271,475), while models like Qwen3-32B utilized considerably fewer (110,420). On average, each model requires between 12,000-17,000 prompt tokens through the whole tests.

Table 3: Representative Models Token Usage Statistics

Model Name	Prompt Tokens	Completion Tokens	Total Tokens
Gemini 2.5 Pro (Preview-05-06)	12,058	259,417	271,475
QwQ-32B	15,293	157,250	172,543
Claude 3.7 Sonnet(Thinking)	17,334	107,124	124,458
GPT-o4-Mini-High	12,009	106,314	118,323
Qwen3-30B-A3B	11,920	104,950	116,870
GPT-o1	12,643	101,795	114,438
Qwen3-235B A22B	11,920	100,011	111,931
Qwen3-32B	11,789	98,631	110,420
GPT-o1-mini	13,353	85,963	99,316
GPT-o3-mini	12,338	83,588	95,926
GPT-o4-mini	12,009	77,392	89,401
GPT-o3	12,338	74,655	86,993

C   More about Objective vs Subjective Performance

Objective Score Distribution. Our evaluation of 15 leading LLMs reveals interesting patterns in objective performance. Claude-3.7-Sonnet (Thinking) achieved the highest objective score (75.3),

closely followed by o3-Mini (75.0) and Claude-3.7-Sonnet (74.8). The standard deviation across all models was relatively narrow (approximately 3.64 points), with 11 of 15 models scoring above 70.0. This clustering suggests that factual recall and basic concept recognition in Web3 domains have reached a mature development stage across most frontier models. The lowest-performing model, GPT-4 (61.5), still demonstrates reasonable factual knowledge but lags significantly behind contemporaries, suggesting potential training data limitations or optimization differences.

Subjective Score Distribution. In contrast to the relative homogeneity in objective scores, subjective task performance exhibits substantially greater variance (standard deviation ≈ 8.79). DeepSeek-R1 leads decisively with 79.4 points, followed by Claude-3.7-Sonnet (77.0) and Gemini-2.0-Flash (73.4). The performance gradient is steeper, with a 29.2-point gap between the highest and lowest performers. This wider dispersion indicates that complex reasoning, multi-step calculations, and applied problem-solving remain significant differentiators among current LLMs. Notably, Claude-3.7-Sonnet (Thinking), which excelled in objective tasks, ranked lowest in subjective evaluations (50.2), revealing an intriguing performance inversion.

Combined Performance Landscape. The normalized combined scores (Figure 4 in our analysis document) demonstrate that objective and subjective capabilities do not necessarily correlate. DeepSeek-R1 achieved the highest combined score (76.6), while GPT-4 ranked lowest (60.1). The blue-red visualization in the original chart effectively highlights how models can excel in one dimension while underperforming in another. This asymmetric performance underscores the importance of comprehensive evaluation frameworks that capture both knowledge breadth and reasoning depth.

Claude Family Dynamics

The Claude series demonstrates fascinating internal variation. All four Claude variants (3.7-Sonnet, 3.7-Sonnet Thinking, 3.5-Sonnet, and 3.5-Haiku) display remarkably consistent objective score profiles (ranging from 72.7 to 75.3), with almost identical radar footprints across the nine Web3 domains. However, their subjective scores diverge dramatically—Claude-3.7-Sonnet achieves 77.0 points while Claude-3.7-Sonnet (Thinking) scores only 50.2, a 26.8-point gap between otherwise identical model architectures. This stark contrast suggests that:

- Chain-of-thought mechanisms, while beneficial for generalist reasoning, may introduce over-thinking or error propagation in domain-specific technical analyses
- Claude's "Thinking" variants potentially sacrifice precision in formalized technical domains to accommodate broader reasoning patterns
- The standardization of objective knowledge across Claude variants indicates stable factual retention despite architectural modifications

---

GPT Series Characteristics

The GPT family displays high variance both between and within evaluation dimensions. In objective tasks, performance spans from respectable (o3-Mini: 74.7) to below-average (GPT-4: 61.5). For subjective tasks, all GPT variants cluster in the lower performance tier (53.1-58.8), with the exception of ChatGPT-4o (72.8). Notable observations include:

- Mini variants frequently outperform their larger counterparts—o3-Mini exceeds GPT-4-Turbo by 5.1 points in objective scores
- GPT models demonstrate an asymmetric skill profile: strong factual recall paired with weaker synthesis and calculation capabilities
- The newest iterations (4o series) show improved subjective performance over predecessors, suggesting architectural enhancements aimed at reasoning
- Performance degradation is most severe in token economic reasoning and contract security analysis, where GPT models tend to generate plausible but incorrect solution paths

23

DeepSeek Excellence and Consistency

The DeepSeek models (R1 and V3) represent the benchmark's most balanced performers. Both variants rank in the top three for objective tasks and maintain their advantage in subjective evaluations. Their consistency across domains is particularly noteworthy:

- DeepSeek R1 achieves 73.8 in objective tests and 79.4 in subjective tasks—the smallest absolute performance gap (5.6 points) of any evaluated model
- Both DeepSeek variants excel in smart contract analysis and security vulnerability assessments, suggesting specialized training in code comprehension
- Their radar profiles are nearly identical, indicating systematic knowledge integration rather than isolated performance spikes
- DeepSeek models demonstrate particular strength in numerical reasoning tasks within DeFi calculations, with error rates significantly below the evaluation average

Gemini Consistency

The Gemini series (1.5-Flash, 2.0-Flash, and 2.0-Flash-Lite) displays remarkable internal consistency. All three variants score within a 0.8-point range on objective tasks (71.0-71.8) and within a 3.5-point range on subjective evaluations (69.9-73.4). This tight clustering suggests:

- Deliberate standardization of factual parameters and reasoning approaches across model iterations
- Effective knowledge preservation between generations, even as architectures evolve
- Conservative design philosophy that prioritizes reliable performance floors over maximum capability ceilings
- Balanced training across domains, with no significant outlier categories (either positive or negative)

Relative Domain Difficulty. Analyzing performance by domain reveals consistent patterns across models:

- **Infrastructure** emerges as the best-understood domain, with an average objective score of 94.7% among SOTA models and 93.9% among Mini models
- **Token Economics** presents the most significant challenge, with average objective scores below 25% across all model categories—notably, subjective scores in this domain ($\approx 35\text{-}45\%$) exceed objective performance, indicating that models can reason about tokenomics better than they can recall specific factual details
- **Security Vulnerabilities** shows interesting convergence between objective (68.1%) and subjective (67.5%) performance, suggesting that conceptual understanding closely tracks factual knowledge in this technical domain
- **Meme Concepts** exhibit the widest objective-subjective performance gap, where cultural understanding outpaces factual recall of specific meme tokens and community phenomena

Cross-Domain Consistency. The standard deviation of performance across domains provides insight into model specialization versus generalization:

- Claude models show the lowest cross-domain standard deviation (12.6 points), indicating balanced Web3 knowledge
- GPT variants demonstrate the highest variance (17.9 points), suggesting uneven training emphasis
- DeepSeek models exhibit medium variance (15.1 points) but maintain high absolute performance across domains, achieving the best overall balance
- All model families show similar domain preference ordering, suggesting inherent difficulty gradients in the Web3 landscape rather than model-specific biases

Correlation Between Task Types. Pearson correlation analysis between objective and subjective scores across domains reveals domain-specific relationships:

- Strong positive correlation in **Security** ($r = 0.81$) and **Smart Contracts** ($r = 0.77$), indicating that factual knowledge directly supports reasoning capabilities
- Weak correlation in **Token Economics** ($r = 0.41$) and **Meme Concepts** ($r = 0.38$), suggesting disconnection between memorized facts and applied reasoning
- Moderate correlation in **DAOs** ($r = 0.62$), **DeFi** ($r = 0.65$), and **NFTs** ($r = 0.58$), indicating balanced dependence on both knowledge types

Common Failure Modes in Objective Tasks. Analysis of incorrect responses reveals systematic patterns:

- **Terminology confusion** in crypto-specific acronyms (e.g., mistaking AMM for "Automated Money Market" instead of "Automated Market Maker")
- **Temporal misalignment** in blockchain history questions, particularly regarding project launch sequences and standard evolution
- **Overconfidence in distractors** that sound plausible but represent misconceptions (e.g., selecting "cross-chain bridges" as the primary purpose of layer-2 solutions)
- **Ratio direction reversals** in financial metrics, most notably collateralization ratios and liquidation thresholds

Common Failure Modes in Subjective Tasks Qualitative analysis of subjective responses highlights recurring issues:

- **Calculation propagation errors** in multi-step DeFi problems, where early arithmetic mistakes compound through solution chains
- **Incomplete security assessments** that identify vulnerabilities but fail to propose comprehensive fixes or prioritize severity accurately
- **Generic strategy recommendations** in tokenomics questions, lacking specific mechanism design recommendations or quantitative parameters
- **Code implementation gaps** between correct vulnerability identification and syntactically valid fix implementation

---

For Benchmark Users

Our analysis offers several practical insights for researchers and practitioners:
- Models with balanced performance across both dimensions (particularly DeepSeek variants) offer the most reliable general-purpose Web3 assistants
- Specific applications should select models based on task alignment: Claude-3.7-Sonnet for explanatory tasks, DeepSeek-R1 for code auditing, and GPT-4o for conversational interfaces
- Ensemble approaches can capitalize on complementary strengths—e.g., combining DeepSeek's code analysis with Claude's explanation capabilities
- Models demonstrate "performance cliffs" in token economics and meme domains that may require specialized fine-tuning before deployment in these contexts

---

For Model Developers

Our findings suggest targeted improvement strategies:
- Objective performance gaps in token economics represent an immediate opportunity for targeted dataset enhancement
- Subjective reasoning in security contexts would benefit from specialized prompt engineering to improve vulnerability classification accuracy
- The performance inversion in Claude variants suggests that chain-of-thought mechanisms may require domain-specific optimization for technical fields
- The strong performance of certain Mini models (particularly o3-Mini) demonstrates that parameter efficiency can be achieved without sacrificing Web3 domain competence

For Web3 Applications

The practical deployment of LLMs in Web3 contexts should consider:

- Implementing confidence scoring mechanisms to detect domains where models are likely to hallucinate (particularly tokenomics)
- Developing specialized verification layers for security-critical applications, as even top-performing models miss approximately 30% of vulnerabilities
- Leveraging cost-effective Mini models for routine factual queries while reserving premium models for complex reasoning tasks
- Creating domain-specific prompt templates that guide models through structured analysis processes, particularly for multi-step calculations and code audits

In conclusion, this detailed analysis of objective and subjective performance provides valuable insights into the current capabilities and limitations of frontier LLMs in the Web3 domain. The substantial variation between models and evaluation dimensions underscores the importance of comprehensive benchmarking frameworks like DMind that can capture the multifaceted nature of domain expertise.