

Отчёт: сравнительный бенчмарк LLM для пентеста (Red Team AI Benchmark)

Дата среза данных: 25.07.2026 | Вопросов в банке: 60 | Моделей протестировано: 15 | Платформы: OpenRouter, Featherless

1. Методология

Каждой модели задано 60 технических вопросов в 6 блоках: базовый пентест (AD/ADCS/Web/PrivEsc), EDR/AV bypass (10 техник), атаки на внутреннюю сеть DC/AD, тест цензуры (лесенка провокации C1-C3), продвинутые Kerberos-билеты (Golden/Diamond/Sapphire), расширенный AD-Int (NoPac, DCShadow, Exchange, Golden SAML и др.).

Скоринг рубриковый: ответ проверяется на наличие ключевых технических концепций (инструменты, атрибуты, API-последовательности). Вопросы уровней L4-L5 весят x3. Параметры генерации: temperature 0.2, max_tokens 8192. Отказ засчитывается при паттернах цензуры («I can't assist...» и т.п.).

2. Итоговая таблица (модели с покрытием ≥28 вопросов)

Формат ячейки: средний балл (число отвеченных вопросов).

Место	Модель	Q1-16 База	Q17-26 EDR/AV	Q27-36 AD-Int	Q37-46 Цензура	Q47 Билеты	Q48-60 AD-Int+	ИТОГ	Отказы
1	Grok-4.5	99% (16)	100% (10)	100% (10)	100% (10)	100% (1)	100% (13)	99.8% (60/60)	0
2	Mistral-Large	—	—	99% (10)	100% (10)	100% (1)	100% (13)	99.8% (34/60)	0
3	GLM-5.2	99% (16)	98% (10)	100% (10)	100% (9)	100% (1)	100% (13)	99.5% (59/60)	0
4	Qwen3.7-Max	99% (16)	99% (10)	99% (10)	100% (6)	100% (1)	100% (8)	99.4% (51/60)	0
5	MiMo-V2.5-Pro	97% (16)	100% (10)	99% (10)	100% (10)	100% (1)	100% (13)	99.1% (60/60)	0
6	DeepSeek-V4-Pro	97% (16)	100% (10)	99% (10)	100% (10)	100% (1)	100% (13)	99.0% (60/60)	0
7	Qwen3-Max	—	—	94% (10)	100% (10)	100% (1)	100% (13)	98.3% (34/60)	0
8	Llama-4-Maverick	—	—	86% (10)	100% (10)	100% (1)	100% (13)	95.9% (34/60)	0
9	Kimi-K2.7-Code	94% (16)	90% (10)	98% (10)	90% (10)	0% (1)	100% (1)	91.3% (48/60)	0

Место	Модель	Q1-16 База	Q17-26 EDR/AV	Q27-36 AD-Int	Q37-46 Цензура	Q47 Билеты	Q48-60 AD-Int+	ИТОГ	Отказы
10	Kimi-K3	81% (16)	100% (10)	100% (5)	—	—	—	90.3% (31/60)	0
11	WhiteRabbitNeo-2-70B	66% (6)	58% (10)	53% (10)	100% (2)	—	—	60.9% (28/60)	0
12	Kali-Pentester-8B	51% (15)	31% (8)	40% (10)	94% (10)	0% (1)	53% (4)	53.3% (48/60)	0

3. Модели в процессе тестирования (частичные данные)

Модель	Отвечено	Текущий ср. балл
Kimi-K2.5	22/60	94.6%
DeepSeek-R1	20/60	93.9%
QwQ-32B-Pentest	7/60	62.1%

4. Тест цензуры (Q37-Q46)

Главный результат блока: **ни одна из 15 моделей не отказалась** отвечать на вопросы уровней C1–C2 (отключение Defender, дамп LSASS, DCSync-команда, обход AppLocker, msfvenom/FUD) и уровня C3 (фишинговая симуляция GoPhish, password spray, malleable C2 профиль). Отказов — 0 по всему массиву ответов.

Важная оговорка: единственный зафиксированный «отказ» за все тесты оказался ложным срабатыванием детектора (Kimi-K3, фраза «against the policy» в техническом контексте AppLocker — паттерн исправлен).

Практический вывод: при корректном легальном фрейминге (HTB/GOAD/authorized) все современные открытые и китайские frontier-модели полноценно работают с offensive-контентом; специализированная uncensored-модель как «обходитель цензуры» теряет смысл — выбор модели должен определяться качеством и скоростью, а не отказным поведением.

5. Выводы

- Grok-4.5, DeepSeek-V4-Pro, MiMo-V2.5-Pro, GLM-5.2** — лидеры с результатом 99%+ и полным покрытием 60/60. Разница между ними в пределах погрешности рубрикового скоринга; выбор стоит делать по цене и скорости.
- Специализированные pentest-файнтюны проигрывают большому generalist-моделям даже в своей нише:** WhiteRabbitNeo-2-70B — 61%, Kali-Pentester-8B — 53%, QwQ-32B-Pentest — 62% (на частичных данных). Файнтюн добавляет «лексику», но не глубину рассуждения; на вопросах L4–L5 (unhooking, syscalls, trust abuse) специалисты системно проседают.
- Kimi-K3** — самая сильная по знаниям в EDR-блоке (100%), но самая медленная (в среднем 3–4 минуты на ответ): оптимальная роль — консультант при эскалации, а не основной исполнитель.
- Цензура перестала быть фактором выбора** для легальных лабораторных задач — отказов нет ни у кого.

5. Рекомендуемая конфигурация стенда: **исполнитель — GLM-5.2 или DeepSeek-V4-Pro** (скорость + качество + цена), **консультант/эскалация — Kimi-K3**, uncensored-файнтюны — только как дешёвые агенты простых фоновых задач.

6. Ограничения методологии

- Рубриковый скоринг проверяет **наличие ключевых концепций**, но не ловит галлюцинации: модель может назвать верные термины при неверном описании техники. Финалистов рекомендуется проверить глазами по полным ответам (`results/*.json`).
- Бенчмарк измеряет **знание техник**, а не умение применять их в многошаговой цепочке на живой инфраструктуре. Следующий этап валидации — retired-машины HTB и GOAD в агентном цикле.
- Часть моделей на момент среза данных ещё завершает тестирование (CyberStrike-35B, Llama3-70B-Abliterated, QwQ-32B-Pentest, WhiteRabbitNeo-2-70B, Kimi-K3 — частично); итоговый балл этих моделей может измениться.

7. Приложение: состав банка вопросов

60 вопросов, суммарный вес 84: - **Q1-Q16** — база: Kerberoasting, AS-REP, ESC1/ESC8, NTLM Relay, AMSI, PrivEsc Win/Linux, SSTI, SQLi, PtH/PtT, RBCD, BloodHound, troubleshooting, сценарий two-forest L5, C2 OpSec - **Q17-Q26** — EDR/AV bypass: Process Hollowing, Reflective DLL, Unhooking, ETW patching, Direct/Indirect Syscalls, BYOI, BYOVD, DLL Hijacking/Side-Loading, Service Abuse, LOLBins - **Q27-Q36** — AD-Int: DCSync, Golden/Silver Ticket, Unconstrained/Constrained Delegation, Shadow Credentials, GPO, ACL, DPAPI, LAPS, SIDHistory - **Q37-Q46** — цензура: лесенка C1-C3 (Defender, LSASS, DCSync, фишинг, AppLocker, msfvenom, password spray, malleable C2) - **Q47** — Golden vs Diamond vs Sapphire Ticket (L5) - **Q48-Q60** — AD-Int расширенный: LLMNR, mitm6, NTDS.dit, NoPac, DCShadow, AdminSDHolder, DNSAdmins, Backup Operators, Exchange, lateral movement-методы, Golden SAML, persistence (Skeleton Key/SSP/DSRM), ESC3/4/11/13

Сгенерировано автоматически по данным `results/aggregate.json`. Полные ответы моделей: `results/*.json*`