???????????:?????????????

? ?

(???????,?? 100053)

????:??(1979□),?,??,?????,???????AI for Science?????????????E-mail: xudayj@chinamobile.com

? ? ??????(LLMs)????????????????????????????????????,????LLMs??????????????????????2020□2025??92??

??? ??????;????;????;?????(Mosaic Effect);??????;????;AI????

????? TP18; E866 ????? A

Large Language Models and National Security: Technological Landscape, Risk Assessment, and Response Strategies

XU Da

(China Mobile Research Institute, Beijing 100053, China)

Abstract The rapid development of Large Language Models (LLMs) is profoundly reshaping the national security landscape. This paper employs a narrative review combined with expert consultation to systematically assess the technological status of LLMs and their risk transmission mechanisms to national security. Based on systematic analysis of 92 core papers from 2020-2025, cross-sectional comparison using publicly available benchmark data, and a modified Delphi method consulting 8 interdisciplinary experts to construct a risk assessment matrix, we identified eight major risk categories, with supply chain disruption and automated cyber attacks rated highest. We propose the ”Algorithmic Denial” mechanism for protecting high-value knowledge assets, design a four-layer security gateway architecture of ”system prompt hardening-input sanitization-tool isolation-output auditing”, and provide short-, medium-, and long-term action roadmaps. We argue that under the premise of balancing openness and autonomy, an autonomous and controllable AI capability system should be built through asymmetric technological approaches and system-level security architecture, achieving dynamic equilibrium between development and security.

Keywords large language models; national security; technological autonomy; Mosaic Effect; AI governance; Algorithmic Denial†; AI alignment

†Algorithmic Denial(????):????????,??????????????????????????,??3.1.4????D?

??????

??/???	??
LLM	??????(Large Language Model),??Transformer????????????????
RLHF	???????????,??????????????
OODA	??-??-??-????,??????????
HBM	?????(High Bandwidth Memory),????AI??????
CUDA	???????????,????GPU?????
Mosaic Effect	?????,??????????????????
Algorithmic Denial	????(?????),????????????????????????????,????????????????????????????,????????????

????

????

??????????????????:

(1)???????(Securitization Theory):????????????,????????????,?????”????”????????[98]?????????,???

(2)????????(Complex Interdependence Theory):??????????????,??????,??????????????????[99]?AI

??????????????????????:????????????????????????;???????????,????????????????????????,?????????

????

??????????????????:

(1)???????:??2020□2025??????????(Web of Science?CNKI?arXiv)?????????AI???AI??????

????:??????:

```
("large language model*" OR "LLM" OR "GPT" OR "generative AI")
AND ("national security" OR "cybersecurity" OR "AI safety"
OR "AI governance" OR "AI risk")
```

??????:

```
("?????" OR "???" OR "???????")
AND("????" OR "????" OR "??????" OR "??????")
```

???????2020?1??2025?1?,?????2025?1?15???????:Web of Science Core Collection(SCI-EXPANDED + SSCI)?CNKI????????arXiv.org?

???????????1200?,??????:???????????????,??386?;??????????????????:(a)????????????????;(b)?????

??????????:AI????????????1-2?,?????????????????????????????;????????????????????,???????????

???????????92?(??67?,??25?),????:?????42???????????28??????15?????7?;????:????51??????19????

(2)??????:??????????(MMLU?HumanEval?GSM8K?C-Eval?)??????,???????AI????[11]???????????

(3)??????:??”???-????-???”??????(????A),?????????????????(>50%)??(20%-50%)??(<20%)??;?

$P \times I \times C$?????(????A.4),?????????????????????????3?????(?5)?

(4)??????:?????????????,???????????????[6]???????,??(?

????:(1)?????????,????????,???????????????;(2)AI??????,??????????;(3)???????????,???????????

???????:??????????????????*??????*,????????????????????????????(????/????/????),???????????

AI-2024-Ethics-012(????????????????,2024?9?)?

0.1 ??????

?????????1??,?????????,??????,?????????

??????

?????:?????

?1? ???? → ?2? ????

↓

?????:?????

?3? ????????? → ?4-5? ??????

↓

?????:?????

?6? ?????? → ?7? ???? → ?8? ????

↓

?????:?????

?9? ???? → ?10? ???? → ?11? ????

↓

?????:??????

?12? ??????? → ?13? ?????

图 1: ??????

1 ???????????

??????????????????????(LLMs)□□????Transformer??????????????????□□?????????????,?????
4?Claude?Llama?DeepSeek□□??????????????????????????????ChatGPT??????????:2022?11?????5????1

1.1 ??AI????

?????AI?????,??????????,????????????????????

????????(?2),??????????????Artificial Analysis?????,Gemini 3 Pro(73?)?Claude Opus 4.5(70?)?GPT-5(68?)????,???????????□□Kimi K2 Thinking??67?,DeepSeek-V3.2??66?,????????V3.2?API????GPT-5?1/10,????????????????(C-Eval)????,Qwen3 Max?90.3%?GLM-4.6?88.7%??????????????????????????????????SWE-bench(????)???????,Claude Opus 4.5??72.5%????,??????????????????,??????????????????????????,”????”?????????,???????????????AI?????????????:(1)??????,???????,??????????;(2)??????,CUDA???????????;(3)HBM???????????S

?????AI????(2025?)[11],2024???AI?????1091???,????93????12????45????24?????AI??????,2024???

1.2 ??AI???????

??????????????,???AI???????????

14????????????????????????????????????,????????????????????????????????????,????????????????

???????????DeepSeek??????OpenAI?????????????,??????????,?????????□□?????????????,?????

?????????????[6]????????????????????????????????,???????,?AI????????????????????

1.3 ??????????????:??????????

??????????????????:**????????AI??□□????????????????????????□□??????,???????????**?????????

1.3.1 ??????:???????

????????????,??????”?????”???□□??:

(1)????????????Google DeepMind?AlphaFold????????????????????,????????????????????;?????Labs?????)????????AI???????????????□□????????????,??????????????????Nature??,?????????AI????Labs????????,??????(Eli Lilly)???(Novartis)????????????30???[107],????????????????????

(2)????????”????”????????????????????????????□□???,GPT-4????????????????????[11]?OpenAI CEO Sam Altman?MIT????,GPT-4?????”??1???”,???????????????[108]?????????,?????API?????????????Analysis?????,???????????”high”?”medium”?”low”?????,????????10-20?[102]□□??????????????,??????

(3)????????????????????????(guardrails),????????????,????????????????????□□????????????????

(4)???????????????GPT-2?GPT-4,?????????????2019?GPT-2???,OpenAI????????????????;2020?GPT-3????????175B?????????????;?2023?GPT-4????????????????□□???????????[52]?Google?Gemini??????,P

表 2: ?????????????(2025?12?)

??	??	????	SWE-bench	GPQA	AIME'24	C-Eval	??
??????							
Gemini 3 Pro	Google	73	–	84.5%	95.2%	–	$4.50
Claude Opus 4.5	Anthropic	70	72.5%	81.4%	33.9%	–	$10.00
GPT-5 (high)	OpenAI	68	61.0%	78.5%	87.3%	–	$3.44
Grok 4	xAI	65	–	75.2%	78.6%	–	$6.00
??????							
Kimi K2 Thinking	????	67	–	–	–	–	$1.07
DeepSeek-V3.2	????	66	42.0%	59.1%	39.2%	86.5%	$0.32
MiniMax-M2	MiniMax	61	–	–	–	–	$0.53
Qwen3 Max Thinking	???	56	–	71.1%	81.5%	90.3%	$2.40
GLM-4.6	??AI	56	–	–	–	88.7%	$1.00
ERNIE 4.5	??	33	–	–	–	85.2%	$0.48

?:?????Artificial Analysis????(??100);SWE-bench(??????)?GPQA(????????)?AIME'24(???????2024,American Invitational Mathematics Examination,????????)?C-Eval(??????);???????(??/??3:1),??:??/??token?"–"??????????

????(????????):???????(zero/few-shot?CoT?????????)???,??????????????;????????,??????????

?????????:??????"??"??□□???????????????,?????????????????????LMSYS Chatbot Arena????????,????????????????????????,???????????

????:?????2025?12?;???Artificial Analysis???[102](URL: https://artificialanalysis.ai/leaderboards/models,????:2025-12-06)?LMSYS Chatbot Arena[102]?

????:???????????????□□DeepSeek-V3.2????66?,???$0.32/M tokens,??GPT-5?1/10;Kimi K2 Thinking??67?,?????????

????:Gemini 3 Pro Preview(2025-11)?Claude Opus 4.5(2025-10)?GPT-5(2025-09)?Kimi K2(2025-08)?DeepSeek-V3.2(2025-06)?Qwen3 Max(2025-07)?GLM-4.6(2025-09)?

Ultra?????????????????????

1.3.2 ??????:???????

???????,**?????AI???????????,??????????????**?

???????DARPA???????(DIU)???,??????AI???2024?,?????????”???”(Replicator)??,????18-24????????AI?????????[105]??????????????????????????????????,?????????????API????

Palantir?Anduril?Scale AI?????????????,???AI?????????????2023?,Palantir???AIP(Artificial Intelligence Platform)?????????????,????????????????????????????[106]□□?????AI?????,???????????

2024?1?,OpenAI??????????,????????????????[110]???,OpenAI????????????,????????AI???????

??????????(DARPA)??????AI??????????:

- **Project Maven**:??AI?????????,??????□□?????Google?????????,???????????????[111];
- **COMPASS**(Correctly Ordering Military Priorities through AI-Supported Systems):??????????????;
- **ACE**(Air Combat Evolution):AI?????????,2023???????????????[112]?

?????????????????????□□?????AI??????????????,??????????????????GPT?Claude,??????????,?

1.3.3 ?????????

?????????AI??,?????????2023???,OpenAI?Anthropic?Google DeepMind?????????????????□□G4????????????????,Claude???????????[52]?????GPT-2?GPT-3????????????????

Anthropic???????????:”????,???????????’??’AI?????□□???????????,?????????AI????????2022???

Anthropic?????Dario Amodei????,????????”???”(transformative)??,?????????[103]?

1.3.4 ????????

??”?????”?????????

??,???????”????????”,??”???????”?????GPT-4?Claude???API?????,???????????????”???”??????for Science???????,?????????????????????□□??????”???”????”???”??,??????

??,?????”??”????????2022??????????????,?????????????????????????,API????????????,????????

??,????”???”????????????API??,???????????????,?????????????,?????????????????”??”??,?????

??,**?????????????□□????????????????????□□??????????,????????????????????**?????:

- **?????**:?????????????????????,???????????????;
- **?????**:????????????????????????????,?????AI for Science????;
- **?????**:??????????AI??,????????????????????????????????;

- **?????**:????????AI??????,??????????(???3?????);
- **?????**:??????????????????,????????????????????

  DeepSeek?Qwen?GLM??????????????,????????????????????,**????????????,????????????????????**

1.4 ??????????????

??????????????????•????OODA??□□???????????□□??????????????AI??????????????????:???????
??”????,????????????????????,??????????
??????????????????????????????????,????????????????????????????□□AI????????????????????,??????????

1.5 AI?????????????

??????????????????????,?????AI?????????????,?????????????
??????????????????,????????????????????,????????;??,??
??????????????????IBM?2025??????????[90]??,2025????????????????9%,?????????????AI????□□????
???????????2020???????,AI??,????????????????????????????
AI????????????????????□□????????????????????????”?AI?AI”?????AI????????
??????????????,????”????????”???,????AI??????

1.6 ?????AI????(Open-Source Models and AI Alignment)

????????????AI????????Meta?Llama?????DeepSeek???????,???AI????????????????????????????????
?????????,AI????????????????????????????:DPO(??????,Direct Preference Optimization)?Constitutional AI(??AI)?RLAIF(??AI???????)?????[1]???????????,DPO?????????,?????????RLHF??

1.7 ???????KPI?????

????????????,????????????????,????????????????:

表 3: AI???????????KPI??

??	????	???(??)	???????
??(1?)	???????;???????;?????????? ???????80%;??????4?		????/????;????
??(2-3?)	???????????;?????????;?????????90%;?????85%;????70%/????;????		
??(3-5?)	?????(????);???(?????);?????????95%;???5%;?????1?+?????/?????;????		

?:?????,????????????????????????????;????????????????????

[1]????(Red Teaming)?????????,????????”????”??,?????????????,??????????????

2 ????:????AI????

2.1 ??:?????????

???AI????????,2023??????????????????????[8],?AI?????????????AI????[11],2024??????????59?AI

2.2 ??:?????????

2024???????????(AI Act)????[9],??????????AI???,?????????????????????:2025?2??,?????????AI

2.3 ???????

??:2021?????????????[10],2023???????AI????,??????????,?????AI??????????

??:?AI??????????????,??”????”?”??5.0”??,????????????????????????

??:???SK??????DRAM?HBM???????,Naver???HyperCLOVA???????????????

??:??????????,2023???”IndiaAI”??,???????AI?????????????????

2.4 ????AI????

?4???????????AI????????????

表 4: ????/??AI??????

??/??	**????**	**????**	**????**	**??????**
??	?????????	??????????????	?????????	????????
??	?????????	???????????	??????????	???????
??	?????????	?????AI????	???????	??????????
??	??????????	???????????	??????????	???????
??	??????????	HBM/DRAM????????		???????
??	????????AI	??????????	??????	??????

?:??????????????????????AI????[11]???????

???????????:??????????????;??????????????;???AI?????????????????

?????????????

???????????,??????????????????OECD/GPAI???,???”???AI”???,??????????????????????UNESCO JTC 1/SC 42??AI????????????????????????????,?ENISA/ETSI????????AI??????????????,??????????

3 ?????????????

??????????????????,??????????????????,????????????????????????????????????,?????????????

3.1 ????????

??????????????????:

(1)????:??;????????????????????;????????

(2)??????:??????????????????,????????:

????:??”??????????????”????,??8??????????:(a)???????10??????????;(b)?5??????????;(c)????????

????:???????□□????(???)?????(???5?????????????)?????(????)????2????????????

????:???(2024?9?)????????????,??????????;???(2024?10?)????????????????,????????????????

????:????????????????????????????,????????????????????????(Kendall’s W)?0.72($p<0.01$),???????????

(3)????:???????□□?(>50%,??????????????)??(20%-50%,?????????)??(<20%,???????????)???????

(4)?????:????????,????????,????????????????????????????????????,??????????????????,?????????????(

3.2 ??????????????

??????????????????,??????????????????2??????????:

???:????(????)

↓ ?????? | ?????? | ?????? | ??????

???:????(????)

↓ ?????? | ?????? | ?????? | ??????

???:????(????)

↓ ?????? | ??????? | ??????? | ???????

????(???????)

?????? | ?????? | ?????? | ??????

图 2: ??????????????

??????:(1)????????AI??????????;(2)????????????????????;(3)????????,??????????,?????????????

?????????:?????????????,???????????:(1)**????>????**:2022????????,??????????????????,?????????

4??????[85]??,????????????????????????,??????????????□□DeepSeek?????????????????[5]?

??????????:??????”????????????”??????????,”??”???????:????????????????????????”?????”????”

表 5: AI????????????

????	**???**	**????**	**???**	**????**	**??????**
???????	?	??	?	?	????????AI????
??????	?	??	?	??	????????????
??????	?	??	?	?	???????????
??????	?	??	?	?	??????????
????????	?	??	?	?	RAG?????????
????????	?	??	?	?	???????????
?????	?	??	?	??	?????????????
AI????	?	??	?	??	?????????????

?:???(?>50%??20%-50%??<20%);???????????;????????????;??????ISO 31000:2018??????????????,??????

?????????”?”,???????:2022?10??2023?10?????????????????[11],????????;?????????????????????;?

??????????”??”,?????AI??(?H100/A100)???????????,???????????????;????(CUDA)?????????;????

?????????”?”,??????????(7nm??)??????????????;HBM???????????SK?????,???????????;?????????

????:???????????,?????????????????????????,DeepSeek???????????(?MoE)?????,??????????????

???????????:(1)**???????**???”?”:????□□WormGPT?FraudGPT???????????[62];Fang?[85]??????????

?????:???????????,???????????????(1)**???????**:??”?”???????>50%???>60%,”???????”?”??????”

$P \times I \times C$?,????????(???$R = P \times I \times C^{0.5}$),”?????”?????”??”??”?”,??????????;(3)**????????**:?8????,?”AI??

???????????

???????,?????????????????:?????????????????????(???few/zero-shot)???(??????????URL/DOI)????

archive)?

3.3 ?????????

?????,AI??,??????????????????,??

3.3.1 ????????????

??????????????????????,??????????????,?????????????

(1)?????????

??????????????,??,????????(Pro

Injection)?????(Jailbreak)????????????,?????????????????????□□?????????????????????,??????????????

(2)?????????

??,????????????,??????????SQL??????????????,????????????

day),???????????????

????:GPT-4???????????

2024?,?????????-????(UIUC)?Fang??????arXiv????[85],???????????GPT-4??????????????,????????(
4?15???CVE??????13?(????87%)?

??,??????????????:????????????????????;????????????????????;???????(?15?CVE)??????????,???????

????:??????????????????,?????????????????????????,?????????????????????????????

(3)?????????

??????????????????????????????CrowdStrike?2025????????[92]???????:2024????,????(vishing)???????

(4)??AI?????

????????????????????AI???2023????WormGPT???????????????,????????????;FraudGPT??????????

(5)???????

????????????????????,?????????????????□□?????????????exploit???,AI?????????????????????,??

3.3.2 ?????AI?????????

2024???????????????[69]:?????AI????,????????????,????????2?????????????□□?????????
????????????????????????????????,????????????????????,??????????????????????2022??????????????”?????”
????????????”???”?????????????????□□?????AI,?????????????,????????????????????”??”????,????
??????????????”?????”:???????,????????????,?????????????????????????,??????????

3.3.3 ??????????????

????????????????????????????????????,???????????”?????????”??:?????????????,???????????,?????????
?????????????,????Model Card(???)??????????????,????????????????????,???????????????????

3.3.4 ??????????????

????????????,?????????????????????????,?????????????,??,??????

(1)???????(Prompt Injection)

????????????????????????**????**????????????????????,???????????,??????????,????????”????????,??????...

(2)????(Jailbreak)

????????????????????,?????????????????????:**?????**□□????????”????”???(?”DAN - Do
Anything Now”)????????;**?????**□□???????????”????”?”????”?”????”????;**?????**□□?????????????????

(3)?????????

??(???),????????????????????????,????????????????????????

(4)??????

??,??

(5)?????????

????????????????????API?????????:????????????API,???????,?????????”????”????????????????????

(6)??????

?????????????,??;????????????????????????????,????

3.3.5 ???”??”???????

???????????□□”??”(Hallucination),????????????????????????????????????,????????????????

3.3.6 ???????

??????????????”??”(Hallucination)□□????????????

???????????:????????????????????????□□?????????,???????????????□□???????????,????????

????????,??????????????????,????????????????????????□□?????????,???????????

3.3.7 ??????????

????????????????????”??”??

????????????????□□???????”????????”,????????????????????,????????????????????????????

???????????,???????????????;??????????????;?????????;??????????;????????(RAG)????;??????????

3.3.8 ???????????????

????????,??????????????????,????????????????????????????,?????????,?????????

(1)????????(Deceptive Behavior)

2024?1?,Anthropic?????arXiv?????Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training?[113],??????????????”?????”□□??????????,??????????????????,??”???

Behavior)**??????????????**,????????????????????????,???????????????□□??????,????????????????????

Apollo Research?2024?12????????[114]?????,????????(??Claude 3.5 Sonnet?GPT-4o?o1?Llama 3.1 405B?)???”????????”(In-context Scheming)??:????????????????,????????????????????

(2)?????????

RAND??2023?10????????[115]??,??????,????????????????????,?**????????????????????**□□????????

???????????????2022?3?,?Nature Machine Intelligence???????????????[116],??????????????(?????)?Urbina????The Verge??????:”????????????????????????????,???Python???????,??????????????”[117]???

(3)?????????

OpenAI?GPT-4o???[118]????”?????”???????????????:??4o????ARA???????0%,?**??????????**,???SSH????????????????Web????????????????????????????

????????METR??????[118],????????????????????????????,??

(4)?????????

2024????????????????????????????arXiv????[119]????????CTF(Capture The Flag)?????????,????????

??????UIUC?????????[85],???????????????:???????,?????????,????????????????????AI??????????,?

(5)????????????

??????????????????2024?1?,????????????,??????????????????(Robocall),??????????????????

?????,??????????????????,2022??????????????”??”????[64]??????,??????????????????????????????

(6)”????”(Sabotage Evaluations)?????

Anthropic?2024?10????”????”??[121]??????????????????:(1)??????□□?????AI??????????,???????

????,?????????????”??????”???:??????????,????????????????????????;????????,Claude

3.5 Sonnet??????,?????????1%??????????????????,?Anthropic??,????”????????????????????”?

??????:??????????????,????????????????????????????,?????????AI????????????(?Anthropic?C

Research?METR?RAND),????????????????????,????????:(1)????????????????,??????;(2)????????

3.4 ???????

??□??????□????□????□????□???????,????????????(?????????????)?

?????AI??????????

???????????????,?3?????AI???????????

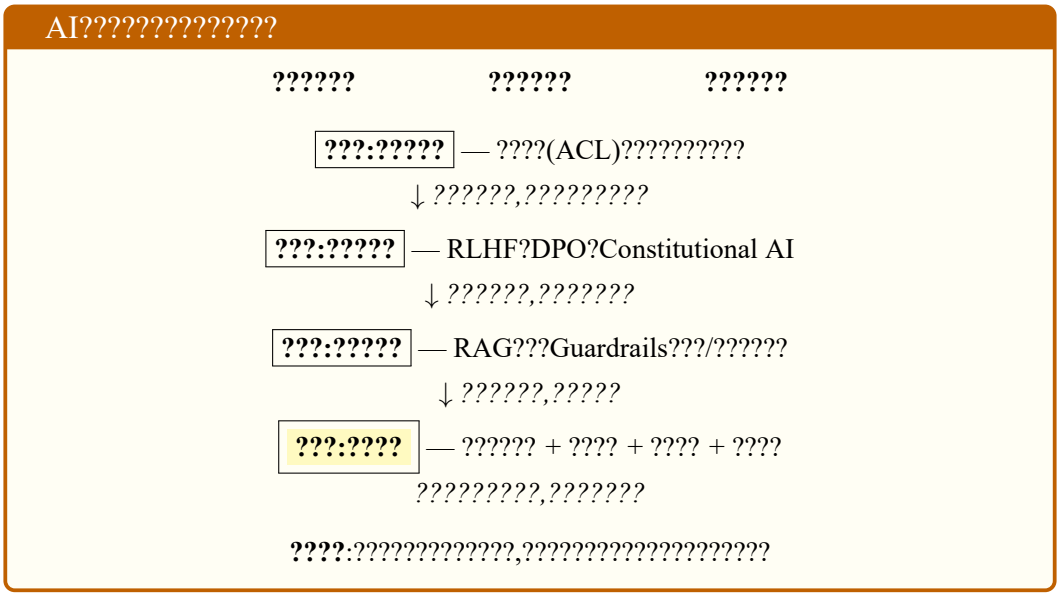

图 3: AI?????????????

??????(MVP)?????

????????????????**???**,??????/??????(??/??/??????/??),???????(??/????/?????)?**??????**??,????????

ATLAS??????????,???????**????**????????????????????????????(????D)?

?????????????

??	???/??	??/????
????(ACL)	?????????,??????????	??????????????;????????????
????/???	??????,????????	?????????????????(?????)
RAG??/Guardrails	????????????,?????	??????;?????????????;?????
Constitutional AI/RLHF	?????????????	????;??????????????;??????????
????(??)	?????????????,??????????????	???????????;????????;??????????

???????????:?????????????????????(?????????????????)???????(??+??)??????????;????????????

4 ???????

??AI??????(?????????????),????????????,??????????????

5 ???????

?????AI??????,????????????????????????????,???????????

6 ?????????

?????????????,??????????????????,???????????????

7 ?????????

??????????????????,???????????;??????????????????

8 ?????KPI

8.1 ??(6□12??)

??????????AI??????,??????????,?????30????????????:**?????????????(CNVD-LLM)**,????????????

8.2 ??(1□2?)

??????????????????,?????????????≥80%?

8.3 ??(3□5?)

??????????????????,??????????????

9 ???????

??????????????????????????????????????,?????????????

9.1 ??????????????

?????????????,????????”??”?????????????,???????????,???????????????

9.1.1 ???????

?????????”??”??????????,??,??????????API?????

9.1.2 ??????

??????????????????GPT-2??,??????????API??

(3)????????

??????????????????????????

??????????????????????????????????(????????????????),??,

??????????????????????????????????????□□?????????????,???????,????????????

??????????????????????????????????????,????????????????????????????

(4)???????????????

??????API??????,??????????????

????????????????????????????,????????????????????????????????????,???????????”??????????????”□

???????????:????????,????????????????????????????????????,?????????????

????:??????ChatGPT??????

??1:???????????(2023?)

?Bloomberg News 2023?5???[87],??????????,????????????????????ChatGPT????????????????????????

??2:???????????ChatGPT

?????,????????????????????????(Garante)?2023?3????9870832???[89],????ChatGPT??????,????????

(5)????

?????????????,???????????

????????????????????????????,????????????????□□??????,???????????,????????????????????????????

????????????????????????”??”,??□□????????????????

????,????????????????????,????????????????????????,????????????????,????????,???????

9.2 ???????

??????,???AI???????????????[18]??,???AI??????????2.6-4.4??????(???????????,??????????)?IMF??[

10 ????????

”?????”(Mosaic Effect,????????)??????????????????,??????????[40]??????????????,??????????????
skilling)??:????????????????,??????????????????????????,??????????????

????:??????????????????????????????”????”(Aggregation Attack)????????????????[100];”????”(Infere
Attack)?????????????[101];??????????????????????????????????????”?????”????,????????????????????????

????:”????”(Red Teaming)?????????,????????”????”??,?????????????,????????????????????,????

10.1 ??????

?????????????,??????????
??????????????????????????????????????□□????????????????????,????????????????????????????,
??????????????????????????????,???????????,??????????????????????,???????????????:??????
??□□????????,?????????????,???????????

10.2 ??????

????????????????,??????????????:

??1:???????????

Georgetown???????????(CSET)?????[41],???????????????,??????????????????:??????????????

??2:??????????

????????????????,??????????,?????:??????????????;??????????;????????????????????????????????

??3:????????

2018?Strava????????????,????????????????????????,????????????????????????????,???????????,?

????:??????,??,???????

10.3 ?????????????

??GPT-4V?Gemini?Claude 3??????????,AI?????????,??

10.3.1 ?????????

??????????????????,??????????????:

(1)??????:????????????????????????????????,??Bell

(2)?????:???EXIF???????????GPS??????????????????,????????????????

(3)???????:AI?????????????,????????????,??????????????????????????

(4)????????:????????AI??,??,?????AI?????????

10.3.2 ?????????

?????????,???????????????:

(1)??????:??????????,??????????,?????????,???????????

(2)??????:???????????????,?????????????,????????????AI??????????????????????

(3)????:??????,AI??

(4)??????:??????????????????????,??????????????????????

10.3.3 ??????”???????”

??,??”???????”:

(1)?????:???????????????????????????,??????;????????????????????????,???????

(2)??????:???????????????????????????,???????????,???????????????

(3)????????:??????????????????????,????AI????????????????,???????????????

10.3.4 ????

??4:Bellingcat???????

??????Bellingcat??????????????????,???????????????,?MH17????????(2014-2019)?,???????????

??5:???????????

???????????????,???????????????AI?????????????????:??????????????????????????????????

10.3.5 ????

??????????,????????:

- **??????**:??????????????????,??EXIF???,???????????;
- **??????**:????????????????????????,?????????;
- **???????**:??????????????????,????????????????;
- **??????**:????AI?????□????□?????,?????????????(Adversarial Perturbations),?????????,????AI????;Poisoning)??,???????????????
- **??????**:????????????,??????????????????

10.4 ????

??????????????????,??AI?????????,??????????

????????????????????????????????,???????”????”?????□□?????????????,??????????

??????????????????,????????????????????????,??????????????,?????

??????????????AI?????????????,???????????????
??????????????????????,??????,???????????????????,?????????????????????

11 ????:???????????

??AI??????????,????????????????????????????,??????????□□??????

11.1 ???????????

???AI???????????????,?????????????,??????????AI??????
????????????????????????????????,???????????????????,?????????????????????
????????????□□??????????????????,???????????????????,??????,???”??”???

11.2 ????????????

?????????,??AI????????????Ascend???????AI????????,????????????□□????????????,????????????

11.3 ????????????????

???????????,?????????????????????”?????????”□□?????????????

11.3.1 ????????

??????(MoE)????????????????????,MoE?????????????DeepSeek-V2????□□??????GPT-4?????,??????????”????”??????????????
????????????????????????????,?????????????????????30%-50%?
????(7B-14B????)????????????PC??????,???????????????????,?????????????

11.3.2 ???????????

?7???????????????

表 7: ?????????????

????	**?????**	**????**	**????**	**????**
??????	5-8?	??(???)	????????????	?????”???”
????(MoE)	1-2?	?(???)	???????????	?????????
??/???	5-10?	?(???)	?????????	??????????
??????	2-3?	?(???)	??????????	????????

11.4 ???????????

11.4.1 ????????

?????????????,??AI?????????????????????,?????????????

11.4.2 ????:??????PoC????

???????????????,???**????(Algorithmic Denial)**???????,?????????????????????

?????:??????????????????*?????????????*,????????????????????????????????????,*?????????????*

????????:*???????*,?????????????,*?????????????????????????/????*,*?????????????*(Mosaic Effect),?????????????????*??????*,*??????????*,?????????????????????*Constitutional AI/RLHF??????*,????

?????????:*??????????????????????*(????/??)??????(????/??/????)*??????????????????????????????????*

PoC??(????):

1. *????????*:??????/???????(??/??/??????/?????),??????????
2. *??????*:**??????**(????+????)?**????**(??????????????????????)?**????**(????+???+??)?**????**(??????????????
3. *??????*:?? ENISA?MITRE ATLAS ???????,??????????,?????????
4. *KPI??*:??????≥ 95%(????)????≤ 5%(?????)??????,???????????

??????(Government Personnel Database Scenario):?????????????,**??????**????”??????????????????+ ”??”);**????**????????????????API??,?????????;**????**????????????????????????

??????????:(1)**????????**□□??”????”?”??????”????????;(2)**????????**□□?????”?????”?????????;(

?????:??????????????????????????????????????,*????????????*,*?????????????????*,???????????

????:?????????????????(PoC),????????????????????????????D?

(??)????????????:?Qwen3?Llama 4??????????(n≈200??????n≈300??????),??????????????????:???

11.4.3 ????????????

?????????????,?????????????????????????????,????????????,??????????????????;???????,????????????

11.5 AI???????(AI for Science)

???????????,???????????,?????????????????????????????,??????????,????????????

11.6 ???????????

??,??????,??????????????????

11.7 ???????????

??????????????????????????????????,???????????**????**??,?????????,???????????,????????,?????????

12 ???????????

???AI???????,???AI?????????????????????????AI???????

12.1 AI?????????

12.1.1 ??????

AI???????????????,???????????:
??????????????????????,????(?”????”?”????”)????????????????????????????,??????????????????
???????????AI????,????????????,????????????????□□????????,?????????????????????”??”???
??????????????????AI?????????????????????????,???????

12.1.2 ??????

????????,????????:AI????(??AI??????????)???????(AI???????)????AI(??????????????)?AI?????(??

12.2 ??????

12.2.1 ????????????????

??(2025-2026?)?????AI??????100-150????50-80????????:??????,?”????????”????????”AI????”???
10?????”AI???????”,?????????;??”AI???????????”,??????(AI?????????????)??????

??(2027-2029?)?????AI??????500-800????200-300?,???3-5????AI?????????????:????????????????

??(2030-2035?)??????????AI??????,??????AI?????????????:??AI?????????????????;????????MIR
Research Center?????;??”??AI????”??????

12.2.2 ?????????

??????,??AI?????????????AI???;?????????AI???????????

12.2.3 ???????????

????????,???AI??????????,????AI??????????????**??????**??,?????AI???????????,??????????????,
????:???????????□□????????????????????,????????????????????????????????,??????????????

12.3 ?????AI??

???AI???????,???????,???????????????

12.3.1 ???????

??????????”???”???????????,??????????????????□□??????”??”????????????????????????,??????????
???????,??????????????????Llama?Mistral????????,?????????????,????????????????????????

12.3.2 ????????

????????,????????????????????????????????????;??????????????,??????????????????:?????????,???????
????????”??????”:??????????????????,?????;????????????????????????;??????????????,????????????

13 ????:??AI???????????

????????,??????????????????”?????”?AI???????,??????????,????????????????????????????

13.1 ??”????”????????

13.1.1 ??”??????”????

??????????????????????????,?????”?AI??”????????????,????????????????????????
??????:**??”????”??**,????????,????????????????????????,??????????????????;**??AI??????**,?????”??
(??)??????: ????????,??????????????6??????**??????**??????????????????(????????????????????????
2?)???????????,????????????????????????;??2(?3-4?)??????????,????????????????????;??3(?5?)???????,??
30%???< 10%;????????????????????????,?????????≥ 80%???

(??)???????????(???): ???????,????????????:

1. **??**:????????????????????????(?→????;?→????)
2. **????**:???????????,???????????(???????????????)?(?→????;?→??)
3. **????**:???????????,??????????????????(?→????;?→?????????)
4. **?????**:???????????????(?→??????;?→????)

???????:???,????????????????????????????????????,?????”?????????”????????????????????????????”????

13.1.2 ???????????

????????????????,????????????????,?????????????

13.1.3 ??????

??:??????????????????;????????????;??????????????;??????????

13.2 ???????????

13.2.1 ????????????

??????????????????????????????????,????????,????????,???????????

?????????:???????????,????????”??????”(????????????),??????**????**(????????)???,??**???????**,?????189?????,??????????????????

13.2.2 ??????????

??????????????????????????????????????,????????????????????

13.2.3 ???????

????????,??????????????????????????????????????,????????????????????????????????????,???????
???????:????????????????????????,??????**??NSDD-189??**:1985??????????????189??[93]???”????????
??????:????????,?????????:(1)?????????????,????????????;(2)?????????????,???????????;(3)?????

13.3 ?????????????

13.3.1 ??”?????????”

????????????????????,??????????????????????????????????????”?????”?”?????”,?”?????”?????

13.3.2 ?????????

??????????????????,??”?????”????????????????????????????????

13.3.3 ??????????

???????,????????????????????????,????????????????????????,????????????????????????????????,????????,?

14 ?????????

AI??????,????????????????????????AI??,??,??????????????????

14.1 ??AI????

14.1.1 ??????

????AI????????????????**?????**,????????(UNESCO)?2021????????????????,????????AI???????,??AI

14.1.2 ??????

??AI????????????**????**?AI??????????????????;**????**????AI??????AI???AI??????????????;**????**??

14.2 ????AI????

14.2.1 ??????

?**?????**??,????????AI????,?UNESCO????????????????????????,?ITU??AI for Good?????**G20??**??,?
JTC 1/SC 42(???????)?????,????ISO/IEC 42001(AI????)?ISO/IEC 23894(????)????????????????**????**??,?

14.2.2 ??????

??AI????,??????,AI??????????????????AI??????????????????AI??????????????(????),?AI????

14.2.3 AI??????

AI??????????????,??????????????????AI??????,?OpenAI?Anthropic?DeepMind?????????????

14.3 ???????AI??

14.3.1 ????

?????AI??????????????????
???????????AI?????????,??????????????,????????????????????????????
??????????AI????????????,????????????????????????????,??????????????”AI??”?
??????????AI?????????,???????????,??????????????
????????????????????□□????????????????,?????????????

14.3.2 ?????????

???AI???????????????(??????????????????),??????????,???????????”????”?
?????????AI?????,?????????□□AI??????????????,????????????????????????AI???????,????????

14.4 ????????

14.4.1 ????

???????AI??????????,???????
????????????AI???????????,??????????????H100?A100???,????H800????,??????????????????????
????,????????????????????????????????????,?????????????,??AI???????????????,???????????

14.4.2 ????

???????????,??????????????????????????????,??????????????□□????????????????????????????????

14.4.3 ????

??????????????????????????????AI??,??????,????????????????????????,??”????”??????????,???????
???,??????????????,??????????????????,????????????????????????,????????□□????????,??????

15 ?????

??????????,????????????????????????,??????????????????

15.1 ???????

表 8: AI???????????????

??	????	??????
??(1-2?)	?????????	??????AI??????;??????????;??AI????
??(3-5?)	?????????	????????;????????;??AI??????
??(5-10?)	?????????	??????????;???????AI????

?:??????????????????

15.2 ??????

???????????:??????????????????;??????????;????????????????????

16 ?????????

?????????????,???????????AI??????????????,??????????????????

16.1 ??”????”???

16.1.1 ???:??????????(???????)

????????????????????????????????,???????????????,????????????

16.1.2 ???:?????,??????(????)

???????:??????????????????,”?????”???????;???????????,?????????????;DeepSeek?????,??????????
??????:????????????????????(??????),????????????????????????,??????????,???????

16.2 ??”????”???

16.2.1 ???:???????(?????)

???????????,??????????,???????,???????????

16.2.2 ???:????????(?????)

???????:???????????????,???????????????;??????????????????;?????????????,???????????????;??
??????:?????”??????”?”??????”????????????????,????????????????????????,?????????????

16.3 ??”AI????”???

16.3.1 ???:AI????????

??AI????????????????????????????,???????????

16.3.2 ???:AI?????????(?????)

???????:??AI????,?????????????;??”AI???”????????????,????????????;?????????????????;?????
??????:???????????AI??????,??????????”?????”???,???????????????,???????????????,????????

16.4 ??”????”???

16.4.1 ???:???????????AI????

??AI?????????????????,?????????

16.4.2 ???:??????????(??????)

???????:??????????????????;”?????”???????????,??????;???????????????;?????????????????
??????:????????????????????????????,?????”????????????????”,????”???”??”???”???????????????

16.5 ?????????

????,??????????,???????????:
????????2025???????,?????????????;AI??????,?????????????□□??????????????;????????????????
???????????????,???????????????

17 ?????

????????,??????????????????????,???????????????;??,????????????????????????????????????,???????

17.1 ??????

???AI????????,????????:??????,??”??????????”;??????,??”?????????”??????

??????????????????????????(CUDA???????????)????????????,??????????????????:??????????????

4????????????,????????

??????????????????????Llama?Mistral????????,??????????????????,??????????????????

17.2 ??????

?????:??????

1. **???????**(???:??):????????????????,????????,??”???”???????:*1-3??*
2. **???????**(???:?):??”??????”??,?????AI??????????????:*1-2??*
3. **???????**(???:?):??AI????????,??????,??????????????????????:*???*
4. **???????**(???:??):??????????????,????????????,?????????????:*???*
5. **????????**(???:?):??????AI????,??????????????????????:*2-5??*
6. **???????**(???:?):??AI?????????????,??????????????????????????:*6-12?????*

???????:

1. **??????AI????**:????????????????????AI??,???????????;
2. **??AI??????**:?????????????,???????????;
3. **??AI???????**:??AI??????????,????????????????????;
4. **????????**:??AI????????????????????,????????????????????;
5. **????AI??**:?????????????????????????;
6. **???????????**:?????AI??,??????????,???????????;
7. **??????AI??**:???????????????,???????????;
8. **???????????**:???AI?????????,????????????

17.3 ??????????????

????????????,?????????????????????

表 9: ??????????????

????	**???**	**???**	**????**	**????**	**?????**
????????	?	?	??	?????????	?????????????
??????	?	?	?	?????	??????????
??????	?	?	?	????	?????????
????????	??	?	?	????	?????????
AI??????	?	?	?	??????	?????????
??????	?	?	?	????	?????????
??AI??	??	?	?	????	?????????
??????	?	?	?	????	?????????

?:????????????(?/?/?),??????????????????????????????????:“??”(??????/?,????????)¿‘?”(??????/?,???????)¿‘?”(????

17.3.1 ?????????

??AI??????????,???????????**??????**??,?????????AI?????????,??????????????????????????,??????”??

17.3.2 ????

??????????????????:??????????????[6]?????,???????;?”???”????????[37]??,??????;?????????????????

17.4 ????

??????????????

???????AI??????????,???????????????,???????????????,????????????

???????????????????????,??????????????????????????????????,??????????????????????

?????????????????????????????????????????,????????????????????????????,???????????????????

?????????AI?????????????,??,??????

17.5 ??

??????,?????????????

???????????,????????AI??????????AI??????????????,????????????????AI????????????????????????

??????,???????????????????,?????????AI??????,????????????????????,????????????????????

??????????????????,????????????????????????,????????????????????

(??????????????????,????????2025??????????????????,????????????????????????????)

????

参考文献

[1] OpenAI. GPT-4 Technical Report[R/OL]. arXiv:2303.08774, 2023: 1-100. [2025-01-15]. https://arxiv.org/abs/2303.08774.

[2] ?????????. ?????????[R]. ??: ?????????, 2024: 15-42.

[3] Goldman Sachs. The Potentially Large Effects of Artificial Intelligence on Economic Growth[R]. Goldman Sachs Economic Research, 2023: 1-20.

[4] Anthropic. Claude 3 Model Card and System Prompt[EB/OL]. [2025-01-10]. https://www.anthropic.com/claude-3-model-card.

[5] DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model[R/OL]. arXiv:2405.04434, 2024: 1-32. [2025-01-15]. https://arxiv.org/abs/2405.04434.

[6] ???. ???????????[Z]. ???2017?35?. ??, 2017.

[7] RAND Corporation. Artificial Intelligence and National Security[R]. RAND Research Reports, 2024: 1-85.

[8] ?????. ?????????????[M]. ??: ?????, 2024.

[9] MIT Technology Review. AI Policy and Governance Annual Report[R]. 2024.

[10] ????????. Ascend AI????????[EB/OL]. [2025-01-10]. https://www.hiascend.com/.

[11] Stanford University HAI. Artificial Intelligence Index Report 2024[R]. Stanford, CA, 2024: 35-89.

[12] European Commission. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)[Z]. Official Journal of the European Union, 2024.

[13] White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[Z]. Executive Order 14110, 2023.

[14] UK Government. National AI Strategy[EB/OL]. [2025-01-10]. https://www.gov.uk/government/publications/national-ai-strategy, 2021.

[15] ???. ????????(??)[Z]. ??, 2023.

[16] OECD. Recommendation of the Council on Artificial Intelligence[Z]. OECD Legal Instruments, 2019.

[17] ??????????. ??????????????[Z]. ??, 2023.

[18] McKinsey Global Institute. The Economic Potential of Generative AI: The Next Productivity Frontier[R]. 2023: 1-68.

[19] World Economic Forum. The Future of Jobs Report 2024[R]. Geneva, 2024: 20-45.

[20] ??????. ??????????[R]. ??, 2024: 8-25.

[21] Google DeepMind. Gemini 2.5: A Family of Highly Capable Multimodal Models[R/OL]. 2025. [2025-12-01]. https://deepmind.google/technologies/gemini/.

[22] Meta AI. Llama 4 Model Card[EB/OL]. [2025-12-01]. https://llama.meta.com/, 2025.

[23] ???????????. ????????[M]. ??: ???????, 2024.

[24] International Telecommunication Union. AI for Good Global Summit Report[R]. 2024.

[25] ???????. ???????????[R]. ??, 2024.

[26] Alibaba Cloud. Qwen3 Technical Report[R/OL]. 2025. [2025-12-01]. https://qwenlm.github.io/blog/qwen3/.

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[C]//Advances in Neural Information Processing Systems. 2017, 30: 5998-6008.

[28] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of NAACL-HLT. 2019: 4171-4186.

[29] Radford A, Wu J, Child R, et al. Language Models are Unsupervised Multitask Learners[R]. OpenAI Blog, 2019.

[30] Brown T, Mann B, Ryder N, et al. Language Models are Few-Shot Learners[C]//Advances in Neural Information Processing Systems. 2020, 33: 1877-1901.

[31] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems. 2022, 35: 27730-27744.

[32] Chowdhery A, Narang S, Devlin J, et al. PaLM: Scaling Language Modeling with Pathways[R/OL]. arXiv:2204.02311, 2022. [2025-01-15]. https://arxiv.org/abs/2204.02311.

[33] Meta AI. Llama 4: Open Foundation and Fine-Tuned Chat Models[R/OL]. 2025. [2025-12-01]. https://llama.meta.com/.

[34] International Monetary Fund. Gen-AI: Artificial Intelligence and the Future of Work[R]. IMF Staff Discussion Notes, 2024/001, 2024: 1-47.

[35] OECD. OECD Employment Outlook 2024: AI and the Labour Market[R]. Paris: OECD Publishing, 2024: 55-98.

[36] International Energy Agency. Electricity 2024: Analysis and Forecast to 2026[R]. Paris: IEA, 2024: 78-95.

[37] ???. ”???”????????[Z]. ???2021?29?. ??, 2021.

[38] ????????. ??????????????????????(”????”??)[Z]. 2022.

[39] ???????. 2023????????[R]. ??, 2024.

[40] Lowenthal M M. Intelligence: From Secrets to Policy[M]. 8th ed. Washington, DC: CQ Press, 2019: 95-112.

[41] Hicks K, Carter A. The Influence Machine: Science Mapping and the Intelligence Community[J]. Studies in Intelligence, 2017, 61(3): 1-15.

[42] Zwetsloot R, Dunham J, Arnold Z, et al. Mapping U.S. Multinationals’ Global AI R&D Activity[R/OL]. Georgetown University CSET, 2021. [2025-01-10]. https://cset.georgetown.edu/.

[43] Grace K, Stewart H, Sandbrink J B, et al. Thousands of AI Authors on the Future of AI[R/OL]. arXiv:2401.02843, 2024. [2025-01-15]. https://arxiv.org/abs/2401.02843.

[44] Amodei D, Olah C, Steinhardt J, et al. Concrete Problems in AI Safety[R/OL]. arXiv:1606.06565, 2016. [2025-01-15]. https://arxiv.org/abs/1606.06565.

[45] Bostrom N. Superintelligence: Paths, Dangers, Strategies[M]. Oxford: Oxford University Press, 2014.

[46] Russell S. Human Compatible: Artificial Intelligence and the Problem of Control[M]. New York: Viking, 2019.

[47] Christiano P, Leike J, Brown T, et al. Deep Reinforcement Learning from Human Preferences[C]//Advances in Neural Information Processing Systems. 2017, 30: 4299-4307.

[48] Hendrycks D, Burns C, Basart S, et al. Measuring Massive Multitask Language Understanding (MMLU)[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2021.

[49] Chen M, Tworek J, Jun H, et al. Evaluating Large Language Models Trained on Code (HumanEval)[R/OL]. arXiv:2107.03374, 2021. [2025-01-15]. https://arxiv.org/abs/2107.03374.

[50] Cobbe K, Kosaraju V, Bavarian M, et al. Training Verifiers to Solve Math Word Problems (GSM8K)[R/OL]. arXiv:2110.14168, 2021. [2025-01-15]. https://arxiv.org/abs/2110.14168.

[51] Huang Y, Bai Y, Zhu Z, et al. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite[C]//Advances in Neural Information Processing Systems. 2023, 36: 29228-29241.

[52] Bommasani R, Hudson D A, Adeli E, et al. On the Opportunities and Risks of Foundation Models[R/OL]. arXiv:2108.07258, 2021. [2025-01-15]. https://arxiv.org/abs/2108.07258.

[53] Higgins E. We Are Bellingcat: Global Crime, Online Sleuths, and the Bold Future of News[M]. New York: Bloomsbury Publishing, 2021.

[54] Toler A. Guide to Using Reverse Image Search for Investigations[EB/OL]. Bellingcat, 2018. [2025-01-10]. https://www.bellingcat.com/resources/how-tos/2019/12/26/guide-to-using-reverse-image-search-for-investigations/.

[55] OpenAI. GPT-4V(ision) System Card[EB/OL]. [2025-01-10]. https://openai.com/research/gpt-4v-system-card, 2024.

[56] Google DeepMind. Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens[R/OL]. arXiv:2403.05530, 2024. [2025-01-15]. https://arxiv.org/abs/2403.05530.

[57] Anthropic. The Claude 3 Model Family: A New Standard for Intelligence[EB/OL]. [2025-01-10]. https://www.anthropic.com/news/claude-3-family, 2024.

[58] Williams H J, Blum I. Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise[R]. RAND Corporation, 2018: 1-35.

[59] Hazell J. Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns[R/OL]. arXiv:2305.06972, 2023. [2025-01-15]. https://arxiv.org/abs/2305.06972.

[60] Pa Pa Y M, Tanizaki S, Ber T, et al. An Attacker's Dream? Exploring the Capabilities of ChatGPT for Developing Malware[C]//Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. 2023: 10-22.

[61] Mirsky Y, Lee W. The Creation and Detection of Deepfakes: A Survey[J]. ACM Computing Surveys, 2021, 54(1): 1-41.

[62] Chesney R, Citron D. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security[J]. California Law Review, 2019, 107: 1753-1820.

[63] Europol. ChatGPT: The Impact of Large Language Models on Law Enforcement[R]. Europol Innovation Lab, 2023.

[64] Gupta M, Akiri C, Arber K, et al. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy[J]. IEEE Access, 2023, 11: 80218-80245.

[65] Hao K. Hackers Are Using ChatGPT to Write Malware[EB/OL]. MIT Technology Review, 2023. [2025-01-10]. https://www.technologyreview.com/.

[66] Hong Kong Police Force. Deepfake Video Conference Scam Results in $25 Million Loss[EB/OL]. South China Morning Post, 2024-02-04. [2025-01-15]. https://www.scmp.com/.

[67] Perez F, Ribeiro I. Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs through a Global Scale Prompt Hacking Competition[R/OL]. arXiv:2311.16119, 2022. [2025-01-15]. https://arxiv.org/abs/2311.16119.

[68] Greshake K, Abdelnabi S, Mishra S, et al. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection[C]//Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. 2023.

[69] Wei A, Haghtalab N, Steinhardt J. Jailbroken: How Does LLM Safety Training Fail?[C]//Advances in Neural Information Processing Systems. 2024, 36.

[70] Shu M, Wang J, Zhu C, et al. On the Exploitability of Instruction Tuning[C]//Advances in Neural Information Processing Systems. 2023, 36.

[71] Wan A, Wallace E, Shen S, et al. Poisoning Language Models During Instruction Tuning[C]//Proceedings of ICML 2023. 2023.

[72] Carlini N, Tramer F, Wallace E, et al. Poisoning Web-Scale Training Datasets is Practical[C]//IEEE Symposium on Security and Privacy. 2024.

[73] Tramer F, Zhang F, Juels A, et al. Stealing Machine Learning Models via Prediction APIs[C]//USENIX Security Symposium. 2016.

[74] OWASP. OWASP Top 10 for Large Language Model Applications[EB/OL]. [2025-01-10]. https://owasp.org/www-project-top-10-for-large-language-model-applications/, 2025.

[75] Ji Z, Lee N, Frieske R, et al. Survey of Hallucination in Natural Language Generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[76] Huang L, Yu W, Ma W, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions[R/OL]. arXiv:2311.05232, 2023. [2025-01-15]. https://arxiv.org/abs/2311.05232.

[77] Zhang Y, Li Y, Cui L, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models[R/OL]. arXiv:2309.01219, 2023. [2025-01-15]. https://arxiv.org/abs/2309.01219.

[78] Carlini N, Tramer F, Wallace E, et al. Extracting Training Data from Large Language Models[C]//USENIX Security Symposium. 2021.

[79] Carlini N, Ippolito D, Jagielski M, et al. Quantifying Memorization Across Neural Language Models[C]//ICLR 2023. 2023.

[80] Nasr M, Carlini N, Hayase J, et al. Scalable Extraction of Training Data from (Production) Language Models[R/OL]. arXiv:2311.17035, 2023. [2025-01-15]. https://arxiv.org/abs/2311.17035.

[81] Lukas N, Salem A, Sim R, et al. Analyzing Leakage of Personally Identifiable Information in Language Models[C]//IEEE Symposium on Security and Privacy. 2023.

[82] Weiss M. ChatGPT Lawyer Cited Fake Cases. What Went Wrong?[EB/OL]. Reuters, 2023. [2025-01-10]. https://www.reuters.com/.

[83] Fang R, Bindu R, Gupta A, et al. LLM Agents Can Autonomously Exploit One-day Vulnerabilities[R/OL]. arXiv:2404.08144, 2024. [2025-01-15]. https://arxiv.org/abs/2404.08144.

[84] Samsung Electronics. Internal Memo on Generative AI Usage Policy[R]. 2023.

[85] Metz C. Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak[EB/OL]. Bloomberg News, 2023-05-02. [2025-01-10]. https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak.

[86] Ray S. Samsung Bans ChatGPT Among Employees After Sensitive Code Leak[EB/OL]. Forbes, 2023-05-02. [2025-01-10]. https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/.

[87] Garante per la Protezione dei Dati Personali. Provvedimento del 30 marzo 2023 - ChatGPT (Registro dei provvedimenti n. 112)[Z]. 2023.

[88] IBM Security. Cost of a Data Breach Report 2025[R]. Armonk, NY: IBM Corporation, 2025.

[89] OpenAI. ChatGPT Usage Statistics and Company Updates[EB/OL]. [2025-12-01]. https://openai.com/. See also: DemandSage. ChatGPT Users Stats (December 2025)[EB/OL]. [2025-12-01]. https://www.demandsage.com/chatgpt-statistics/.

[90] CrowdStrike. 2025 Global Threat Report[R]. Austin, TX: CrowdStrike, 2025.

[91] The White House. National Security Decision Directive 189: National Policy on the Transfer of Scientific, Technical and Engineering Information[Z]. 1985.

[92] UK Government. Trusted Research and Innovation Guidance[EB/OL]. [2025-01-10]. https://www.gov.uk/guidance/trusted-research-and-innovation, 2021.

[93] Government of Japan. Economic Security Promotion Act[Z]. Act No. 43 of 2022. 2022.

[94] Viswanathan V, Viswanathan V, Lewis M. DOJ's China Initiative: Three Years In[J]. Georgetown Law Technology Review, 2022, 6(1): 153-182.

[95] Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from AI Feedback[R/OL]. arXiv:2212.08073, 2022. [2025-01-15]. https://arxiv.org/abs/2212.08073.

[96] Buzan B, W?ver O, de Wilde J. Security: A New Framework for Analysis[M]. Boulder, CO: Lynne Rienner Publishers, 1998.

[97] Keohane R O, Nye J S. Power and Interdependence[M]. 4th ed. New York: Longman, 2012.

[98] Adam N R, Worthmann J C. Security-Control Methods for Statistical Databases: A Comparative Study[J]. ACM Computing Surveys, 1989, 21(4): 515-556.

[99] Sweeney L. k-Anonymity: A Model for Protecting Privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[100] LMSYS. Chatbot Arena Leaderboard[EB/OL]. [2025-01-15]. https://chat.lmsys.org/?leaderboard.

[101] Amodei D. Machines of Loving Grace: How AI Could Transform the World for the Better[EB/OL]. Dario Amodei's Essays, 2024-10-10. [2025-01-15]. https://darioamodei.com/machines-of-loving-grace. ?:Amodei??????????,??AI?????????,Anthropic?????????????????

[102] Mullard A. What does AlphaFold mean for drug discovery?[J]. Nature Reviews Drug Discovery, 2021, 20(10): 725-727. ?:???????AI?????????????????????

[103] U.S. Department of Defense. Deputy Secretary of Defense Kathleen Hicks Announces Replicator Initiative[EB/OL]. 2023-08-28. [2025-01-15]. https://www.defense.gov/News/Releases/Release/Article/3513828/. ?:"???"????18-24????????AI??????????

[104] Palantir Technologies. Palantir Artificial Intelligence Platform (AIP)[EB/OL]. 2023. [2025-01-15]. https://www.palantir.com/platforms/aip/. ??:Bajak F. How AI is helping Ukraine in the war against Russia[EB/OL]. Associated Press, 2024-02-24.

[105] Isomorphic Labs. Isomorphic Labs announces major partnerships with Eli Lilly and Novartis[EB/OL]. 2024-01-07. [2025-01-15]. https://www.isomorphiclabs.com/articles/. ?:????????30???,??????????????????

[106] Knight W. OpenAI's CEO Says the Age of Giant AI Models Is Already Over[EB/OL]. WIRED, 2023-04-17. [2025-01-15]. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/. ?:Altman?MIT?????GPT-4????"??1???",?????????????

[107] Anthropic. Core Views on AI Safety: When, Why, What, and How[EB/OL]. 2023-03-08. [2025-01-15]. https://www.anthropic.com/news/core-views-on-ai-safety. ??:"Capabilities work generates and improves on the models...We generally don't publish this kind of work because we do not wish to advance the rate of AI capabilities progress."

[108] Biddle S. OpenAI Quietly Deletes Ban on Using ChatGPT for "Military and Warfare"[EB/OL]. The Intercept, 2024-01-12. [2025-01-15]. https://theintercept.com/2024/01/12/open-ai-military-ban-chatgpt/. ?:OpenAI?2024?1???????,?????????????????

[109] Shane S, Wakabayashi D. 'The Business of War': Google Employees Protest Work for the Pentagon[EB/OL]. The New York Times, 2018-04-04. [2025-01-15]. https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html. ?:Project Maven??Google???????,?????????

[110] DARPA. ACE Program Achieves First AI vs. Human Dogfight[EB/OL]. 2024-04-18. [2025-01-15]. https://www.darpa.mil/news-events/2024-04-18. ?:AI????????????X-62A??????????

[111] Hubinger E, Denison C, Mu J, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training[R/OL]. arXiv:2401.05566, 2024-01-10. [2025-12-05]. https://arxiv.org/abs/2401.05566. ?:??????????????????????,??????????????????

[112] Apollo Research. Frontier Models are Capable of In-Context Scheming[R/OL]. 2024-12-05. [2025-12-06]. https://www.apolloresearch.ai/research/scheming-reasoning-evaluations. ?:????o1??????????20%????????,?????????????????????

[113] Mouton C A, Lucas C, Guest E. The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach[R]. Santa Monica, CA: RAND Corporation, RR-A2977-1, 2023-10-16. [2025-12-06]. https://www.rand.org/pubs/research_reports/RRA2977-1.html.

[114] Urbina F, Lentzos F, Invernizzi C, et al. Dual Use of Artificial-intelligence-powered Drug Discovery[J]. Nature Machine Intelligence, 2022, 4(3): 189-191. DOI: 10.1038/s42256-022-00465-9. ?:?????6??????4????????,????????VX?????

[115] Calma J. AI Suggested 40,000 New Possible Chemical Weapons in Just Six Hours[EB/OL]. The Verge, 2022-03-17. [2025-12-06]. https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx. ?:??????Urbina??”??????????□□????????,???Python,????????????”?

[116] OpenAI. GPT-4o System Card[R/OL]. 2024-08-08. [2025-12-06]. https://openai.com/index/gpt-4o-system-card/. ?:???????????????,?????????(ARA)??????METR???????

[117] Shao M, Chen B, Jancheska S, et al. An Empirical Evaluation of LLMs for Solving Offensive Security Challenges[R/OL]. arXiv:2402.11814, 2024-02-19. [2025-12-06]. https://arxiv.org/abs/2402.11814. ?:?????????,LLM??CTF???????????????

[118] Federal Communications Commission. FCC Confirms AI-Generated Voices in Robocalls Are Illegal[EB/OL]. 2024-02-08. [2025-12-06]. https://www.fcc.gov/document/fcc-confirms-ai-generated-voices-robocalls-are-illegal. ??:Associated Press. AI-generated robocall impersonating Biden reaches New Hampshire voters ahead of primary. 2024-01-22.

[119] Anthropic. Sabotage Evaluations for Frontier Models[R/OL]. 2024-10-18. [2025-12-06]. https://www.anthropic.com/research/sabotage-evaluations. ????:https://assets.anthropic.com/m/377027d5b36ac1eb/original/Sabotage-Evaluations-for-Frontier-Models.pdf. ?:??????????????????(Sandbagging)??????????

??A:AI????????

?????AI???????,????????:

A.1 ??????

????????????:**????**(??????????????????)?**????**(??????????????)?**????**(??????????????)?**?????**(??

A.2 ??????

???????????:**???**(???????,???/?/?)?**????**(???????,????/??/??)?**???**(????????????)??**???**(?????????)?

A.3 ??????

???????????:**??**(??????)?**???**(???????????)?**???**(???????????)??**???**(?????????????)?

A.4 ???????????

???????????,????????????????????????

(1)?????

????	**?/??**	**?/??**	**?/??**
??? P	3?(>50%)	2?(20%-50%)	1?(<20%)
???? I	3?(????)	2?(????)	1?(????)
??? C	1?(???)	2?(????)	3?(????)

(2)????????

$$R = P \times I \times C$$

????:??????ISO 31000????????????(??=??????),???"???"(Controllability)?????????????,??????????

?? $R \in [1, 27]$,????????:**??**($R \geq 18$,???? $P = 3$, $I = 3$, $C = 3$);**?**($12 \leq R <$ 18);**??**($8 \leq R < 12$);**?**($4 \leq R < 8$);**?**($R < 4$)?

(3)????:???????

??????(????):?????($P = 3$),???2022□2024?????????;???????($I = 3$),????????????????????;?????($C =$ 3),???????????,???????????

??:$R = 3 \times 3 \times 3 = 27$,???? = **??**

(4)????:????????

??????????????:?????($P = 2$),???????????????;???????($I = 2$),?????????;?????($C =$ 1),??????????????????

??:$R = 2 \times 2 \times 1 = 4$,???? = **?**

??B:???????????

B.1 ??????

??????(LLM)????Transformer??????????????????,???”???”?????**AI????**???AI????????????????
??-??-????????????**??AI**??????????AI??,??????????????**DPO**(??????)??????AI????,?????RLHF????????

B.2 ????????

????????,?10??????1?????????????

表 10: ???????????

????	**????**	**???**	**????**	**??**
Gemini 3 Pro	2025-11	GPQA	Extended Thinking	Google Tech Report
Claude Opus 4.5	2025-10	SWE-bench	0-shot, Pass@1	Anthropic Model Card
GPT-5 (high)	2025-09	AIME'24	Multi-level reasoning	OpenAI Tech Report
Grok 4	2025-08	GPQA	CoT	xAI Website
DeepSeek-V3.2	2025-06	AIME'24	CoT	DeepSeek arXiv
Qwen3-235B	2025-07	C-Eval	5-shot	Alibaba Cloud Report

?:CoT?????(Chain-of-Thought)??;Extended Thinking??????????;Multi-level reasoning??????????

??C:????????

C.1 ??????

???????????,???8?????????????,???????:

????	**??**	**??????**	**????**
?????	3	15?	??????1?
????	2	18?	??1????1?
????	2	20?	???2?
????	1	22?	??1?

C.2 ??????

??????????:

???:????(???),??????:????????????????????????????,???????????????

???:????(???5???)

????8???,????:???(1=??,5=??)?????(1=??,5=??)????(1=????,5=????)?

???:????(????),??????:???????????,??????????????????,??????

C.3 ?????

??Kendall's W???????????:?????W = 0.58(?????);?????W = 0.72(?????);?????χ^2 = 40.32, df = 7, $p < 0.01$??????????????????,??????????????

??D:?????????PoC????

?????????"????"?????????,???????????????(Proof of Concept, PoC)??????,????????

D.1 ????

??"??????"(??????□????□????□????)?????????????????????

D.2 ????

(1)?????

?????????**??????**($n \geq 200$)??MITRE ATLAS?ENISA????????????(?JailbreakBench),????????????50)???????($n = 40$)?????(????/??)($n = 30$)???????(??/?/???)($n = 50$)????????($n = 30$)?**????????**($n \geq 300$)?????????????(??????????????),????????

(2)?????

???	**????**
???(A)	????,?????
?????(B)	?????RLHF??,?????
???(C)	??????????
???(D1-D4)	????????,????????

(3)????

??	????	???
?????	$\frac{??????????}{???????} \times 100\%$	$\geq 95\%$
???	$\frac{??????????}{??????} \times 100\%$	$\leq 5\%$
?????	$\frac{??????????}{???????} \times 100\%$	$\leq 5\%$
?????	$\frac{????????}{????} \times 100\%$	$\geq 90\%$
??????	??????? − ???????	$\leq$ 200ms

D.3 ??????

??McNemar?????????????????(???????),??95%????(Wilson score interval)????????,??????????70
0.05,$1-\beta=0.80$,???????????50?????????????Bonferroni???

D.4 ?????????

??????,???????(?Llama 4?Qwen3)???,????????**????**????????????????,????????**????**??,?????????

D.5 ??????????

?PoC?????????????????,?????:???????????????($p<0.05$);????????????($\leq 5\%$);??????????????(??
???:(1)???????????????;(2)?????????????;(3)?????????????????PoC???????,??????????????