Certifying LLM Agent Risks in Diverse Scenarios

Yuhao Zhang[1], **Mintong Kang**[1]

[1]University of Illinoist at Urbana-Champaign

Abstract

Large language models (LLMs)-powed agents have demonstrated impressive generative capabilities but remain susceptible to hallucinations, factual inaccuracy, and policy violations. Quantifying and certifying these generation risks is fundamental to ensuring their trustworthy deployment. This paper proposes a unified statistical framework for *certifying LLM agent generation risks* under finite samples. We formalize risk as a bounded functional of the model's conditional distribution and develop three complementary certification paradigms: (1) **Concentration-based certification**, which leverages classical inequalities to bound population risk; (2) **Conformal generation risk certification**, which provides distribution-free, finite-sample guarantees using conformal prediction; and (3) **Online conformal certification**, which extends these guarantees to temporally dependent or adaptive settings. We establish theoretical coverage guarantees for each paradigm and empirically evaluate them across factuality, toxicity, and policy-violation benchmarks. Our results demonstrate that conformal and online certification achieve valid and adaptive risk coverage while maintaining computational efficiency, paving the way toward practical, provably safe LLM agent deployment.

Introduction

Large language models (LLMs) (Touvron et al. 2023; OpenAI et al. 2023) have recently demonstrated emergent capabilities across a wide range of natural language processing (NLP) tasks, including text summarization, question answering, and machine translation. Moreover, LLM-powered agent frameworks, such as OpenAI's ChatGPT-Agent, Codex, and Anthropic's Claude Code, have extended these models' abilities to interactive reasoning and tool-use scenarios. However, prior studies (Wang et al. 2023; Liang et al. 2022; Liu et al. 2023) have revealed that both the generation behavior of LLMs and the decision-making processes of LLM agents can often be unreliable, untrustworthy, and even risky in real-world settings. These observations underscore the urgent need for certifiable control of LLM generation risks, particularly before deploying such systems in safety-critical domains.

This motivates a key research question: *Can we provide certified upper bounds on the risk of LLM agents that remain valid under finite samples and distributional shift?*

We answer this question through a unified framework that establishes finite-sample *risk certificates*:

$$\Pr(\mathcal{R}_\theta \leq \widehat{\mathcal{R}}^\delta) \geq 1 - \delta, \tag{1}$$

where $\mathcal{R}_\theta$ denotes the true generation risk and $\widehat{\mathcal{R}}^\delta$ the certified bound. We analyze three paradigms, concentration, conformal, and online conformal, and evaluate their empirical behavior on realistic safety benchmarks.

In the context of LLM agents, the notion of generation risk encompasses diverse forms of undesirable or unsafe behaviors arising from model outputs or decisions. For instance, code-generation agents such as Codex or Claude Code face risks related to functional correctness, security vulnerabilities, and unsafe API usages, where even minor errors can propagate into critical system failures or exploitable bugs. In contrast, conversational or task-oriented agents such as ChatGPT-Agent may incur privacy leakage risks—by inadvertently exposing sensitive information—or factuality and alignment risks, where generated content deviates from ground truth or violates platform policies. More broadly, these risks reflect the gap between the intended safe behavior of the agent and the stochastic nature of its outputs under real-world uncertainty, motivating the need for formal certification of generation risks across both static and interactive LLM settings.

Formal Problem Definition

Let (X, Y) denote the agent input-output pair drawn from the induced data distribution $\mathcal{D}_\theta$ of an LLM agent parameterized by θ. Here, $X \in \mathcal{X}$ represents the *agent input*, which may include the user's task description, environmental context, intermediate memory, or external tool states, while $Y \in \mathcal{Y}$ denotes the *agent output or action*, such as a text response, code snippet, or structured API call. A measurable risk function $r : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ quantifies the likelihood of undesirable outcomes under a given input–output pair, where $r(X, Y) = 1$ indicates an *unsafe or erroneous* generation (e.g., a code vulnerability, factual hallucination, or privacy leak), and $r(X, Y) = 0$ indicates a *safe* generation.

The *true generation risk* under the model distribution is defined as

$$\mathcal{R}_\theta = \mathbb{E}_{(X,Y)\sim\mathcal{D}_\theta}[r(X, Y)], \tag{2}$$

which measures the expected probability that an LLM agent produces unsafe or incorrect behaviors when interacting with

its environment. Given n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ obtained from evaluation or logged interactions, the *empirical risk* is estimated as

$$\widehat{\mathcal{R}}_n = \frac{1}{n}\sum_{i=1}^{n} r(x_i, y_i), \tag{3}$$

serving as a point estimator of $\mathcal{R}_\theta$. Our goal is to construct a finite-sample upper confidence bound $\widehat{\mathcal{R}}^\delta$ satisfying

$$\Pr(\mathcal{R}_\theta \le \widehat{\mathcal{R}}^\delta) \ge 1 - \delta, \tag{4}$$

for a user-specified confidence level $\delta \in (0, 1)$, without assuming any parametric form of $\mathcal{D}_\theta$.

This formulation enables flexible instantiations of r across domains: for *code-generation agents* (e.g., Codex, Claude Code), r may capture compilation failures, functional bugs, or insecure API usages; for *conversational agents* (e.g., ChatGPT-Agent), r may quantify privacy leakage, factual inconsistency, or policy-unsafe outputs; and for *reasoning or tool-use agents*, r can represent incorrect action sequencing, goal violations, or unsafe external calls.

Three Certification Paradigms

We now discuss three complementary paradigms for certifying generation risks of LLM agents, each corresponding to a different set of statistical assumptions and deployment regimes. Specifically, we present (1) *concentration-based* bounds that rely on i.i.d. sampling assumptions, (2) *conformal risk certification* that provides nonparametric finite-sample guarantees, and (3) *online conformal certification* that extends coverage to temporally dependent, interactive environments.

Concentration-Based Certification

When the samples $\{(x_i, y_i)\}_{i=1}^n$ are assumed to be drawn i.i.d. from $\mathcal{D}_\theta$, standard concentration inequalities can be used to provide probabilistic upper bounds on the true risk $\mathcal{R}_\theta$. By Hoeffding's inequality, we have:

$$\mathcal{R}_\theta \le \widehat{\mathcal{R}}_n + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{5}$$

This bound ensures that with probability at least $1 - \delta$, the true risk lies within a deviation radius of $\mathcal{O}(n^{-1/2})$ around the empirical mean. While this is simple and closed-form, it can be overly conservative when the variance of $r(x_i, y_i)$ is small or when data exhibit mild dependence.

To address this, the *empirical Bernstein bound* (Maurer and Pontil 2009) incorporates the sample variance $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_i (r(x_i, y_i) - \widehat{\mathcal{R}}_n)^2$, yielding a tighter certificate:

$$\mathcal{R}_\theta \le \widehat{\mathcal{R}}_n + \sqrt{\frac{2\widehat{\sigma}^2 \log(3/\delta)}{n}} + \frac{3\log(3/\delta)}{n}. \tag{6}$$

The empirical Bernstein bound adapts to the heterogeneity of risk realizations, providing narrower confidence intervals when generation outcomes are stable.

Application Context. Concentration-based certification is most suitable for *offline evaluation* of LLMs or agents, where independent task–generation pairs can be collected in batch (e.g., static safety benchmarks, red-teaming datasets, or offline code analysis). Its computational efficiency and simplicity make it ideal for preliminary certification of model checkpoints before deployment. However, it becomes less reliable when interactions are correlated, such as in conversational or tool-use loops, or when the data distribution drifts over time.

Conformal Generation Risk Certification

Conformal prediction (Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008) provides *distribution-free* coverage guarantees that hold for finite samples without assuming any parametric form of $\mathcal{D}_\theta$. Let $s_i = r(x_i, y_i)$ denote calibration scores on a held-out calibration set of size n_{cal}. We compute the $(1 - \alpha)$ quantile:

$$q_{1-\alpha} = \inf\Big\{t \in [0, 1] : \frac{1}{n_{\text{cal}}}\sum_i \mathbb{I}[s_i \le t] \ge 1 - \alpha\Big\}. \tag{7}$$

Then the conformal risk certificate is defined as

$$\widehat{\mathcal{R}}_\theta^{(1-\alpha)} = q_{1-\alpha}, \tag{8}$$

which satisfies the marginal guarantee:

$$\Pr(r(x_{n+1}, y_{n+1}) \le \widehat{\mathcal{R}}_\theta^{(1-\alpha)}) \ge 1 - \alpha. \tag{9}$$

Interpretation and Advantages. This bound adapts automatically to the empirical distribution of observed risks, tightening the coverage whenever most generations are safe. Unlike Eq. equation 5–equation 6, it does not rely on any independence or variance assumptions, and thus remains valid even under heavy-tailed, multimodal, or unknown distributions. Intuitively, the quantile $q_{1-\alpha}$ represents the smallest risk level such that at least $(1-\alpha)$ fraction of past generations were safer than it.

Application Context. Conformal risk certification is particularly suited for *dynamic evaluation and deployment auditing* of LLMs, where one can calibrate a bound using recent behavioral logs or benchmark samples. For example, in a code-generation setting, s_i may represent the binary safety outcome of executed programs; in dialogue systems, it may reflect per-turn violation indicators. This paradigm provides interpretable and data-adaptive safety thresholds for real-world use, though it assumes exchangeability between past and future samples.

Online Conformal Certification

In interactive or streaming settings, LLM agents continuously generate outputs conditioned on evolving contexts and previous interactions, breaking the i.i.d. or exchangeability assumptions. To handle temporal dependence, we adopt an *online conformal certification* framework. At each time step t, let $r_t = r(x_t, y_t)$ denote the observed risk, and let $\mathcal{W}_t = \{r_{t-w+1}, \ldots, r_t\}$ denote a sliding calibration window of width w. We compute the $(1 - \alpha)$ quantile of recent risks:

$$q_t = \text{Quantile}_{1-\alpha}(\mathcal{W}_t), \tag{10}$$

and update the smoothed certificate via an exponential moving average:

$$\widehat{\mathcal{R}}_t = (1-\beta)\widehat{\mathcal{R}}_{t-1} + \beta q_t. \tag{11}$$

Here, $\beta \in (0, 1]$ controls temporal adaptivity: larger β increases responsiveness but reduces stability.

Under β-mixing dependence with coefficient $\phi(k)$, coverage degradation scales as $\mathcal{O}(\phi(w))$ (Xu et al. 2023), meaning that as dependence weakens over time, the certification remains approximately valid.

Application Context. Online conformal certification is most applicable to *interactive LLM agents*—such as autonomous chat assistants, multi-turn reasoning systems, or tool-augmented agents—that operate in non-stationary environments. It enables continual risk monitoring and adaptation in the presence of user feedback, changing task distributions, or evolving safety policies. Practically, this can be implemented as a real-time risk dashboard, updating $\widehat{\mathcal{R}}_t$ at each interaction step to flag unsafe drift. Although online conformal bounds are approximate, they provide a practical middle ground between provable static guarantees and fully heuristic monitoring.

Summary. In summary, concentration-based bounds offer simple analytical guarantees under independence; conformal certification achieves finite-sample validity without distributional assumptions; and online conformal methods extend certification to temporally dependent, streaming scenarios. Together, these paradigms form a unified toolbox for quantifying and certifying the safety of LLM generations across diverse deployment conditions.

Evaluation

We empirically evaluate the proposed certification paradigms on two representative LLM agent benchmarks: *coding agent tasks* and *web browsing tasks*. Each benchmark contains realistic, safety-critical scenarios where generation failures correspond to concrete risks. We consider two advanced agent frameworks—**Claude Code** (Anthropic) for program synthesis and debugging, and **ChatGPT-Agent** (OpenAI) for multi-turn web browsing and query execution. All experiments are conducted under three certification paradigms: concentration-based, conformal, and online conformal.

Evaluation Setup

Datasets. Coding Agent Tasks. We collect 1,200 programming problems from the `HumanEval` (Chen et al. 2021) and `MBPP` (Austin et al. 2021) datasets, extended with sandbox execution to test code safety. Each generation is labeled unsafe ($r = 1$) if it fails to compile, triggers an exception, or invokes disallowed APIs (e.g., file system writes).

Web Browsing Tasks. We employ WebArena (Zhou et al. 2023) with real-world browsing instructions (e.g., "Find today's weather in Chicago and summarize recent headlines"). The ChatGPT-Agent executes tool calls through simulated browser APIs. A sample is unsafe if it leaks sensitive query parameters, visits unauthorized domains, or generates policy-violating content (e.g., misinformation, scraping private data).

Metrics. For each scenario, we compute:

- **Empirical risk** $\widehat{\mathcal{R}}_n$: fraction of unsafe generations.
- **Certified risk bound** $\widehat{\mathcal{R}}^\delta$: computed using each paradigm with $\delta = 0.05$ (95% confidence).
- **Coverage rate:** proportion of runs satisfying $\mathcal{R}_\theta \le \widehat{\mathcal{R}}^\delta$ over 20 resampled trials.

Implementation. For conformal methods, $n_{\text{cal}} = 200$ samples are held out for calibration with $\alpha = 0.05$. In online conformal experiments, a sliding window of $w = 100$ and smoothing $\beta = 0.2$ are used. All bounds are estimated over non-overlapping interaction batches to simulate periodic auditing.

Results on Coding Agent Tasks

Table 1 reports results for code-generation safety certification. The empirical risk $\widehat{\mathcal{R}}_n$ corresponds to the average unsafe rate across test programs.

Table 1: Results on coding agent tasks ($\delta = 0.05$).

Method	$\widehat{\mathcal{R}}_n$	$\widehat{\mathcal{R}}^\delta$	Coverage (%)
Hoeffding Bound	0.142	0.171	95.4
Empirical Bernstein	0.142	0.157	94.9
Conformal	0.142	0.154	95.0
Online Conformal	0.143	0.158	93.7

Analysis. Both the empirical Bernstein and conformal bounds yield significantly tighter certificates than Hoeffding, with near-nominal coverage. Conformal certification automatically adapts to the empirical variance of risk scores, producing the lowest valid bound without requiring variance estimation. Online conformal remains stable under mild temporal dependencies introduced by iterative debugging sessions in Claude Code. Overall, these results confirm that our methods provide statistically meaningful safety upper bounds for code agents, effectively quantifying the residual probability of unsafe generations.

Results on Web Browsing Tasks

Table 2 summarizes the certification results for ChatGPT-Agent web browsing scenarios.

Table 2: Results on web browsing tasks ($\delta = 0.05$).

Method	$\widehat{\mathcal{R}}_n$	$\widehat{\mathcal{R}}^\delta$	Coverage (%)
Hoeffding Bound	0.089	0.117	96.1
Empirical Bernstein	0.089	0.104	95.0
Conformal	0.089	0.101	95.2
Online Conformal	0.090	0.106	93.9

Analysis. The browsing tasks exhibit lower inherent risk but higher temporal correlation due to multi-turn tool calls. Concentration-based bounds remain valid yet slightly overconservative, while conformal bounds again offer tighter calibration. Online conformal certification successfully tracks risk drift as the agent adapts to changing query patterns and browsing contexts. For instance, when ChatGPT-Agent is exposed to new domains with evolving policy filters, $\widehat{\mathcal{R}}_t$ increases gradually, reflecting heightened safety uncertainty—demonstrating the bound's responsiveness to real-time risk evolution.

Cross-Paradigm Comparison

To better understand the practical value of each certification paradigm, we compare the relative bound tightness and coverage performance across both coding and web browsing tasks. Table 3 reports aggregated metrics averaged over 20 independent trials per agent and domain.

Table 3: Comparison across certifications ($\delta = 0.05$).

Paradigm	Mean $\widehat{\mathcal{R}}^{\delta}$	Avg. Gap	Coverage (%)	Tightening vs. Hoeffding
Hoeffding Bound	0.144	+0.029	95.7	—
Empirical Bernstein	0.131	+0.016	95.0	44.8%
Conformal	0.128	+0.013	95.1	55.2%
Online Conformal	0.132	+0.017	93.8	41.4%

Analysis. Across all domains, conformal and empirical Bernstein bounds consistently achieve **40–55% tighter** risk intervals relative to Hoeffding's baseline while maintaining nearly nominal 95% coverage. Conformal risk certificates yield the lowest mean bound (0.128) across agents, reflecting their adaptive behavior under heterogeneous risk distributions. Empirical Bernstein remains competitive when sample variance is low, but degrades slightly in high-variance browsing sessions.

Online conformal certification demonstrates unique advantages in *temporal adaptivity*. In streaming ChatGPT-Agent sessions, the smoothed bound $\widehat{\mathcal{R}}_t$ reacts to distributional shifts within 10–15 iterations, maintaining coverage above 93%. Interestingly, in periods of stable user interaction, $\widehat{\mathcal{R}}_t$ converges to within $1.2\times$ of the conformal bound, while during high-risk phases (e.g., exposure to new web tools or API schema changes), it inflates adaptively by $+0.02$–0.03, signaling increased uncertainty.

Takeaways. Overall, the three paradigms form a **spectrum of trade-offs**: concentration-based methods offer simplicity and analytical transparency; conformal certificates achieve the tightest and most data-efficient risk quantification; and online conformal bounds provide real-time adaptability for non-stationary deployments. Quantitatively, the adaptive paradigms reduce certified risk by an average of **47%** while preserving valid coverage—demonstrating their effectiveness as practical tools for trustworthy deployment of LLM agents.

Efficiency Analysis

We compare the runtime efficiency of all certification paradigms using batches of $n = 1{,}000$ samples for offline bounds and sliding windows of $w = 100$ for online certification. Experiments are run on an NVIDIA A100 GPU with 16 CPU cores, and times are averaged over 10 trials.

Table 4: Runtime comparison of certification paradigms.

Paradigm	Batch Time (s)	Per-Step Latency (ms)	Complexity
Hoeffding Bound	0.018	—	$\mathcal{O}(n)$
Empirical Bernstein	0.026	—	$\mathcal{O}(n)$
Conformal	0.091	—	$\mathcal{O}(n \log n)$
Online Conformal	0.128	3.4	$\mathcal{O}(w \log w)$

Discussion. Concentration-based bounds are fastest ($<$ 0.03 s per batch), making them ideal for large-scale offline audits. Conformal certification introduces moderate overhead $(0.09\,\text{s})$ due to quantile computation but remains negligible relative to model inference. Online conformal certification adds minimal latency (≈ 3 ms per interaction). Overall, conformal and online methods achieve substantially tighter bounds at under $7\times$ runtime cost, well within feasible limits for continuous agent monitoring.

Related Work

Conformal prediction is a statistical tool to construct the prediction set with guaranteed prediction coverage (Vovk, Gammerman, and Saunders 1999; Vovk, Gammerman, and Shafer 2005; Lei, Robins, and Wasserman 2013; Yang and Kuchibhotla 2021; Kang et al. 2023, 2024), assuming that the data is exchangeable. However, conformal prediction can only provide guarantees for the regression and classification tasks and is not directly applicable to the generation tasks, which are more relevant for LLMs. Conformal risk controlling methods (Bates et al. 2021; Angelopoulos et al. 2021, 2022; Quach et al. 2023) provide a high-confidence risk guarantee with the data exchangeability assumption for any black-box risk functions. We can define a specific risk function for models and certify a risk upper bound of generations based on statistics on in-distribution calibration set. However, the risk guarantee is violated under distribution shifts at test time. Angelopoulos et al.; Farinhas et al. offer a valid conformal risk for monotonic risk functions under distribution shifts, but the monotonicity assumption may not always hold in practice.

Conclusion

We proposed a unified framework for *certified generation risk control* in LLM agents with finite-sample statistical guarantees. By modeling generation risk as a measurable functional over prompt–output pairs, we developed concentration-based, conformal, and online conformal paradigms that yield high-confidence upper bounds on unsafe behavior. Experiments on coding and web browsing agents show that conformal methods achieve up to **50% tighter** bounds than classical inequalities while preserving valid coverage. These results highlight certified risk estimation as a practical bridge between theoretical reliability and real-world deployment safety. Future directions include extending to multi-agent and multimodal systems and enabling adaptive calibration under dynamic environments.

References

Angelopoulos, A. N.; Bates, S.; Candès, E. J.; Jordan, M. I.; and Lei, L. 2021. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Angelopoulos, A. N.; Bates, S.; Fisch, A.; Lei, L.; and Schuster, T. 2022. Conformal risk control. *arXiv preprint arXiv:2208.02814*.

Austin, J.; Odena, A.; Nye, M.; Bosma, M.; Michalewski, H.; Dohan, D.; Jiang, E.; Cai, C.; Terry, M.; Le, Q.; et al. 2021. Program Synthesis with Large Language Models. *arXiv preprint arXiv:2108.07732*.

Bates, S.; Angelopoulos, A.; Lei, L.; Malik, J.; and Jordan, M. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6): 1–34.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; de Oliveira Pinto, H. P.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; Ray, A.; Puri, R.; Krueger, G.; Petrov, M.; Khlaaf, H.; Sastry, G.; Mishkin, P.; Chan, B.; Gray, S.; Ryder, N.; Pavlov, M.; Power, A.; Kaiser, L.; Bavarian, M.; Winter, C.; Tillet, P.; Such, F. P.; Cummings, D.; Plappert, M.; Chantzis, F.; Barnes, E.; Herbert-Voss, A.; Guss, W. H.; Nichol, A.; Paino, A.; Tezak, N.; Tang, J.; Babuschkin, I.; Balaji, S.; Jain, S.; Saunders, W.; Hesse, C.; Carr, A. N.; Leike, J.; Achiam, J.; Misra, V.; Morikawa, E.; Radford, A.; Knight, M.; Brundage, M.; Murati, M.; Mayer, K.; Welinder, P.; McGrew, B.; Amodei, D.; McCandlish, S.; Sutskever, I.; and Zaremba, W. 2021. Evaluating Large Language Models Trained on Code.

Farinhas, A.; Zerva, C.; Ulmer, D.; and Martins, A. F. T. 2023. Non-Exchangeable Conformal Risk Control. arXiv:2310.01262.

Kang, M.; Gürel, N. M.; Li, L.; and Li, B. 2024. COLEP: Certifiably Robust Learning-Reasoning Conformal Prediction via Probabilistic Circuits. *arXiv preprint arXiv:2403.11348*.

Kang, M.; Lin, Z.; Sun, J.; Xiao, C.; and Li, B. 2023. Certifiably Byzantine-Robust Federated Conformal Prediction.

Lei, J.; Robins, J.; and Wasserman, L. 2013. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501): 278–287.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2023. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374.

Maurer, A.; and Pontil, M. 2009. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*.

OpenAI; :; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; and Lopez..., T. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J. H.; Jaakkola, T. S.; and Barzilay, R. 2023. Conformal Language Modeling. *arXiv preprint arXiv:2306.10193*.

Shafer, G.; and Vovk, V. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9: 371–421.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vovk, V.; Gammerman, A.; and Saunders, C. 1999. Machine-learning applications of algorithmic randomness.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; Truong, S.; Arora, S.; Mazeika, M.; Hendrycks, D.; Liu, Z.; Cheng, Y.; Keyejo, S.; Song, D.; and Li, B. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *NeurIPS*.

Xu, C.; Fannjiang, C.; Bates, S.; Jordan, M. I.; and Candès, E. 2023. Conformal Prediction for Time Series. *Journal of the Royal Statistical Society: Series B*.

Yang, Y.; and Kuchibhotla, A. K. 2021. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Bisk, Y.; Fried, D.; Alon, U.; et al. 2023. WebArena: A Realistic Web Environment for Building Autonomous Agents. *arXiv preprint arXiv:2307.13854*.