Condensing Law of Security Information

Qi Tan[*,1], Tianyu Cui[*,2], Qi Li[✉,3], and Ke Xu[✉,4]

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Independent researcher
[3]Institute for Network Sciences and Cyberspace, Tsinghua University
[4]Department of Computer Science and Technology, Tsinghua University

Abstract

Scaling laws indicate that the generalization error of a large language model (LLM) decreases monotonically with model size once data is sufficient, and the corollary that larger models outperform smaller ones is now widely accepted. We argue that this corollary fails in security: decisions there hinge on anomalous inputs, which are long-tailed (e.g., zero-day attacks) and therefore poorly served by likelihood fitting. Worse still, the capability brought about by scale (e.g., inferring the intended whole from a corrupted part) leads a LLM silently *repair* the anomaly before any downstream security head ever sees it. We formalize this as a tradeoff between two quantities: the cognitive capability $K(W) = I(W;P)$ that a model holds about the world distribution, and the security-information retention $R(W) = \frac{I(Z;S)}{I(A;S)}$ that survives in its intermediate representation. We derive three information-theoretic bounds to connect them: an upper bound on restoration accuracy in terms of $K(W)$, an upper bound on $R(W)$ in terms of the restoration error, and a Fano lower bound on security-task error in terms of $R(W)$. These theorems approximately imply that usable security information is a product of how much security information survives and how well it can be extracted, indicating that optimal usable security information can be achieved at a certain model size. We state this as the *condensing law of security information*: usable security information is unimodal in model size, so security generalization error is U-shaped and is minimized at a finite $N^\star$ rather than decreasing without bound. The experiments on various real-world LLMs validate the condensing law, indicating that there is an optimal model size to solve security issues.

Index Terms—large language models, scaling laws, condensing law, security, representation learning

1 Introduction

Artificial intelligence, and in particular the large language model (LLM), has rapidly advanced in recent years. Now it plays a substantial role in domains where public welfare depends:

[*]These authors contributed equally to this work.

health, transport, finance, and etc. From the release of ChatGPT in 2022, which marked the arrival of LLMs in public view, to present-day that LLMs can locate security vulnerabilities and solve open problems in mathematics, the capability of these models is no longer confined to language, they now act at the leading edge of the physical world.

The prevailing view is that, given enough training data, the final performance of an LLM improves steadily with model size. That view produced the scaling laws [1, 2], which state that the generalization error is a function of both data volume and model size:

$$L(N;D) \;=\; E \;+\; \frac{A}{N^{\alpha}} \;+\; \frac{B}{D^{\beta}}. \tag{1}$$

Here E is the irreducible error of the language model, a property of the intrinsic uncertainty of language itself, N is the model size and D is the size of the training data. Moreover, A, B, α and β are four constants capturing the linear and exponential growth coefficients in the model size and data size, respectively.

Specifically, Eq. (1) decomposes pre-training error into three terms: the irreducible error, the error due to limited model size, and the error that caused by limited data. In 2022, DeepMind fitted the four unknown coefficients empirically, training over 400 models with sizes from 70 million to 16 billion parameters on corpora from 5 to 500 billion tokens [2]. With the fitted curve, they posed a constrained optimization problem and solved for the best ratio of data size to model size under a fixed computation budget, providing a principled allocation rule for subsequent LLM training.

Scaling laws tell us that spending more data on more parameters yields a better model, and the trajectory of the field has broadly borne the prediction out: larger parameter counts have brought stronger cognitive capability.

Subsequently, as deployment matured, efficiency became a second axis of interest. Following the method of scaling laws, a group from Tsinghua University proposed the *densing law* [3], an empirical fit of the marginal capability gain per parameter, intended to guide the design of on-device models. It takes the form

$$\ln \rho_{\max} \;=\; At + B, \tag{2}$$

where $\rho_{\max}$ is the capability density of a model, estimated as the ratio between the smallest parameter count that attains the same error after distillation and the model's actual size. By estimating the density of widely used open-source base models released at different dates, the authors fitted A and B on 29 open-source base models, finding that capability density has risen steadily with scale, doubling approximately every 3.3 months.

For a long time, "larger is better" has been the consensus. Yet an empirical regularity, however useful as guidance, may fail to explain certain domains. Security is one, and there scaling laws must be supplemented to offer real insight.

2 Rethinking "Larger Is Better" in Security

In security, the consensus is contested from two directions at once. On the one hand, scaling laws tell us that a larger model has stronger cognitive capability and can handle more complex tasks. On the other hand, large models specialize at likelihood fitting, and are by construction

insensitive to the tail of the distribution. Yet the tail is precisely where security information lives. These two trends run against each other, so the "larger is better" consensus cannot be carried over to security unexamined.

Simply put, security tasks depend on information carried by anomalous data, yet anomalies are rare compared with normal traffic. Being long-tailed (e.g., zero-day attacks), they are easily overlooked by LLMs, which hurts security-task performance. More sharply, as the model size grows, their picture of the world becomes more complete and their ability to infer intent from partial input improves accordingly, as in code completion and text continuation. Reconstructing a whole from a part helps on mainstream tasks that depend on regularities of the world. On security tasks, however, it has the opposite effect: the model uses what it knows about normal content to *correct* the anomaly that carries the security signal, so that less of that signal survives into whatever downstream component has to make the security judgment.

A small model lacks the cognitive capability to extract enough task information to complete a security judgment. A large model repairs the anomalous data and so reduces the security information retained downstream. This tradeoff motivates the central conjecture of this paper.

Conjecture 1 (Condensing law of security information)**.** *In the security domain, there exists an optimal model size matched to the security task, at which security information attains a condensed form.*

Section 5 gives the functional form of the Condensing Law once the two quantities it trades off have been defined.

3 Problem Formulation

To explain why security is not a "larger is better" domain, and to gain a deep insight about the mechanism rather than the phenomenon, we analyze the flow of security information directly. In particular, we use the following notation.

- P: the world distribution, which is the distribution of normal data.
- $\mathscr{D}$: the pre-training data, $\mathscr{D} \sim P$.
- W: the model parameters, $W = \mathscr{F}(\mathscr{D}; U)$, where U collects randomness unrelated to training.

These three variables form a Markov chain $P \to \mathscr{D} \to W$. The pre-training data is drawn from the world distribution, and the data is the model's only channel to the world distribution.

For a test sample in the security setting, we define:

- X: the normal content of the test input, $X \sim P$; $|\Lambda|$ is the size of its alphabet (i.e. the number of values X may take).
- Δ: the anomalous content, introduced by an attack (by obfuscation or by injecting an error).
- A: the observed input, $A = \mathscr{H}(X; \Delta)$, indicating that the observed data is determined jointly by the normal and the anomalous content.

- S: the security label, determined by the anomalous content, $S = \psi(\Delta)$; $|\mathscr{S}|$ is the size of its alphabet.
- Z: the intermediate representation the model extracts.

To advance the theoretical analysis, we make two simplifying assumptions.

Assumption 1 (Fresh sample)**.** $W \perp (X, \Delta, A) \mid P$, which indicates that the test sample does not appear in the pre-training data.

Assumption 2 (Label independent of normal content)**.** $S \perp X$, indicating whether the input is anomalous is unrelated to the normal content it carries.

With this notation, we reduce the conjecture to a quantitative statement about two parameter-dependent quantities.

- *Cognitive capability*, $K(W) = I(W;P)$: the information about the world distribution that the model holds.
- *Security-information retention*, $R(W) = \frac{I(Z;S)}{I(A;S)}$: the fraction of the security-relevant information that survives the model's feature extraction.

4 Theoretical Analysis

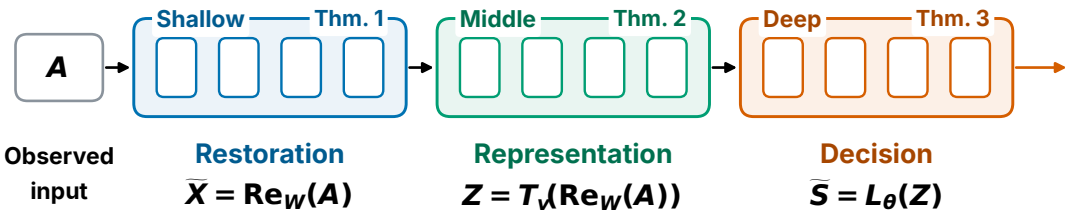


Figure 1: Pipeline of LLM

As displayed in Fig. 1, we divide the pipeline of LLM inference into three stages: (i) the restoration stage, where the shallow layers aim to repair the attacked input so that the output follows their own perception of the real-world; (ii) the representation extraction stage, where the middle layers extract features from the fundamental knowledge, which has been repaired to a familiar format, for completing downstream tasks; (iii) the decision stage, where LLMs provides the predictions by analyzing the extracted features.

Based on the decomposition, we derive the following theorems, which motivates us to propose the Condensing Law of Security Information.

Theorem 1 (Upper bound on correction ability)**.** *Let $\widetilde{X} = \mathrm{Re}_W(A)$ be the model's estimate of the clean content X from the attacked observation A under parameters W, and let $\varepsilon = P(X \neq \widetilde{X})$ be its error rate. Then*

$$1 - \varepsilon \;\leq\; B_1 + A_1 K(W), \tag{3}$$

where $A_1 = \frac{1}{\log|\Lambda|}$ and $B_1 = 1 - \frac{H(X|A)-1}{\log|\Lambda|}$.

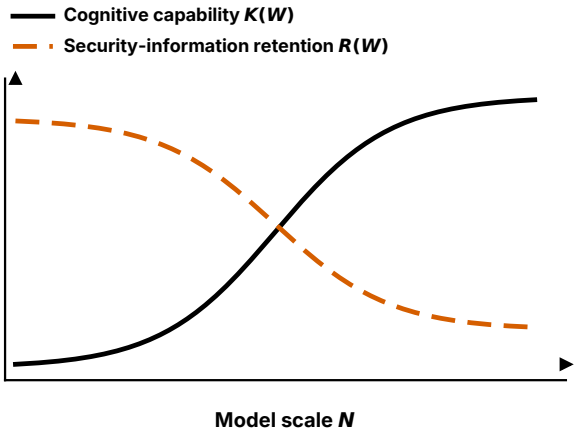


Figure 2: Cognitive capability and security-information retention versus model size. Cognitive capability is upper-bounded by $H(P)$. By Theorem 2, retention decreases affinely with capability, so against model size it is a reflected (and rescaled) copy of the capability curve.

In Theorem 1, $1-\varepsilon$ is the probability that the model's correction restores the normal content. The larger it is, the stronger the model's correction ability. The theorem indicates that when cognitive capability is insufficient (i.e., $K(W)$ is small), the model can correct attacked input only to a limited degree and its correction ability $1-\varepsilon$ stays small, which implies strong correction ability *requires* strong cognitive capability.

For the retention of security information, we have the following theorem.

Theorem 2 (Upper bound on retention)**.** *If $Z = T_V\left[\mathrm{Re}_W(A)\right]$, which indicates that Z depends on A only through the intermediate variable $\widetilde{X}$, then the retention of attack-related security information satisfies*

$$R(W) \;\le\; C_1\varepsilon + D_1, \tag{4}$$

where $C_1 = \frac{\log|\Lambda|}{I(A;S)}$ and $D_1 = \frac{1}{I(A;S)}$.

Theorem 2 demonstrates that when the model corrects anomalies well (i.e., ε is small), the share of security information retained in its output shrinks.

Finally, we prove the following theorem.

Theorem 3 (Lower bound on security-task error)**.** *Predicting S from the intermediate representation Z, with $\widetilde{S} = L_\theta(Z)$,*

$$P(\widetilde{S} \neq S) \;\ge\; L_1 - M_1 R(W), \tag{5}$$

where $L_1 = \frac{H(S)-1}{\log|\mathscr{S}|}$ and $M_1 = \frac{I(A;S)}{\log|\mathscr{S}|}$.

Theorem 3 states that as the retained security information decreases, the model's ability to recognize attack-induced security problems weakens and the error rate rises sharply.

These theorems show that the security information retained in intermediate representations decreases as the model's cognitive capability grows, which motivates the Condensing Law in Section 5.

5 The Condensing Law

The analysis in Section 4 yields the picture in Fig. 2. Specifically, scaling laws indicate that as model size grows, the generalization error decreases and the model's knowledge of the world grows with it. Cognitive capability, in turn, is positively correlated with correction ability, and as correction ability grows, it steadily removes security information from the features passed to the higher layers.

To handle a security task efficiently, a model must be good at *acquiring* security information, which depends on two things: how much security information is present at all, and how well the model can extract it. The first is the quantity of raw material, and the second is the ability to refine it. The joint behavior of the two curves in Fig. 2 is therefore what determines the security information a model can actually use.

Because these two move in opposite directions with scale, their combination gives the *usable security information*

$$U(N) \;=\; K\big(W(N)\big)\cdot R\big(W(N)\big). \tag{6}$$

As shown in Fig. 3, the product form is an ansatz, not a consequence of Section 4. In particular, refining ability is of no use without raw material (i.e., the security information), and raw material is of no use without the ability to refine it. Composing U with the relation in Theorem 3 then gives a U-shaped generalization-error curve over model size.

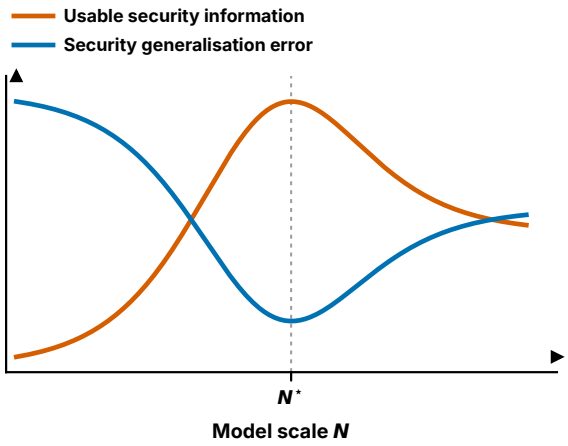


Figure 3: The Condensing Law: usable security information $U(N)$ is the product of the two curves of Fig. 2 and peaks at an intermediate model size $N^{\star}$. Security generalization error is minimized at the same point.

Conjecture 2 (Condensing law)**.** *$K(W(N))$ is non-decreasing in N and bounded above by $H(P)$; $R(W(N))$ is non-increasing in N by Theorem 2; and $U(N)$ in* (6) *is unimodal. Consequently there is a finite*

$$N^{\star} \;=\; \arg\max_{N} U(N) \tag{7}$$

at which the security generalization error attains its minimum, and beyond which further scale makes security performance worse rather than better.

Loosely speaking, as $K(W) \geq \frac{1-B_1}{A_1} - \frac{1}{A_1}\varepsilon$ and $R(W) \leq C_1\varepsilon + D_1$, we have:

$$U(N) = K(W) \cdot R(W) \approx \left[\frac{1-B_1}{A_1} - \frac{1}{A_1}\varepsilon\right] \cdot [C_1\varepsilon + D_1], \tag{8}$$

where ε is a function of N. This equation approximately holds because we cross multiply the two sides of the inequalities.

Eq. (8) is a quadratic function of ε, therefore has a unique maximum point in terms to ε. That is,

$$\frac{dU(N)}{d\varepsilon} = \frac{-2C_1}{A_1}\varepsilon - \frac{D_1}{A_1} + \frac{(1-B_1)C_1}{A_1} = 0 \Rightarrow \varepsilon^* = \frac{1-B_1}{2} - \frac{D_1}{2C_1}, \tag{9}$$

Hence, if $\varepsilon^* = \frac{H(X|A)-2}{\log|\Lambda|}$, where $\varepsilon^* \in (0,1)$ is reasonable, the usable security information achieves its optimum

$$U^*(N) = \frac{H(X|A)^2}{4 \cdot I(A;S)}. \tag{10}$$

N^* is defined as the solution of Eq. (10), which maximizes the usable security information to accomplish the downstream missions.

At $N^\star$, cognitive capability and the retention of security information are best traded off. Hence, security information condenses there, and the model's ability to solve security problems is maximized.

6 Experiments

We conduct two experiments to validate the conjecture. The first tests the mechanism we propose, and the second tests the consequence the conjecture predicts.

6.1 The Retention of Security Information Decreases with Model Size

Dataset. In the first experiment, the benign and anomalous inputs in our comparison are taken from *HttpParamsDataset* [4], whose records are the *values* of HTTP request parameters. It holds 31,067 items: 19,304 benign and 11,763 anomalous, the latter split into SQL injection (10,852), cross-site scripting (532), path traversal (290) and command injection (89). In particular, the benign items are the payloads of the normal requests in HTTP DATASET CSIC 2010 [5], automatically generated from a Spanish-language e-commerce application. The SQL injections were produced with `sqlmap` [6], the cross-site scripting samples are generated with XSSYA and FuzzDB [7], and the command-injection and path-traversal samples are produced with Vega Scanner and FuzzDB.

In this experiment, we aim to validate that as the size of a model increases, the intermediate-layer representations produced by normal and anomalous inputs move closer together. This result implies that security-information retention is decreasing, validating Theorem 2.

To achieve this target, we measure the distance between the centroids of the attacked and the normal representations at matched absolute block indices across the Pythia suite [8], whose members share one corpus in one order [9] and therefore isolate scale from training-data confounds. To analyze these results in detail, we display the distances of all attacks separately

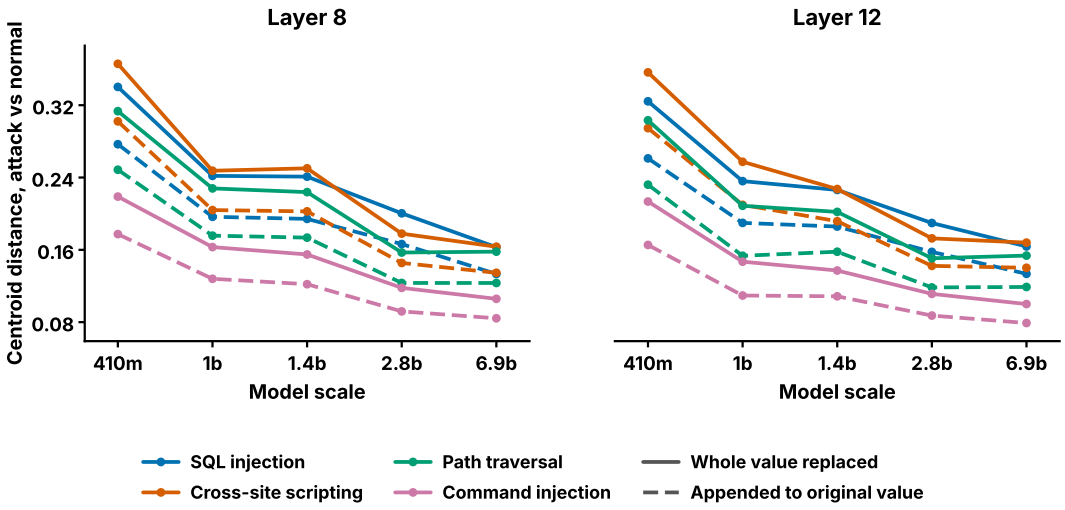


Figure 4: Centroid distance between attacked and normal representations at blocks 8 and 12, with the eight attack conditions kept separate. Every condition falls with model size.

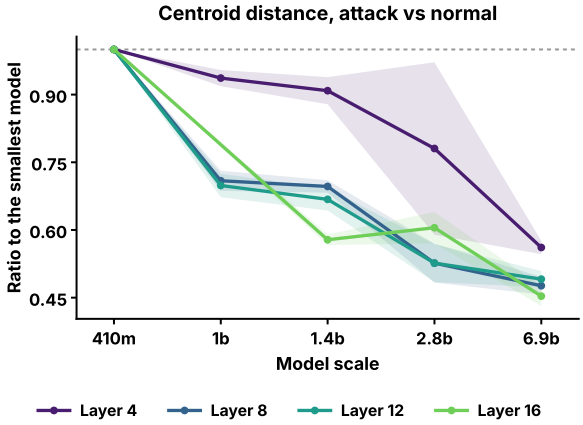


Figure 5: The same measurements normalized by each condition's value at the smallest model and averaged over the eight conditions.

in Fig. 4. The distance of all attacks declines with scale, which validates that the retained security information is decreasing. Moreover, Fig. 5 normalizes each attacks by its own value at the smallest model, which removes the level differences between classes and leaves the consistency of the trend visible.

6.2 Real-World Fitting

As size grows, performance on the security task has a best fit at an intermediate point, confirming the condensing law conjecture. We use the cybersecurity world model benchmark in BigModelGym [10] to evaluate the real-world fitting across existing LLMs. The benchmark dataset contains 1,000 security-related data chains extracted from code (250), traffic data (250), system logs (250), and agent trajectories (250) in real-world systems.

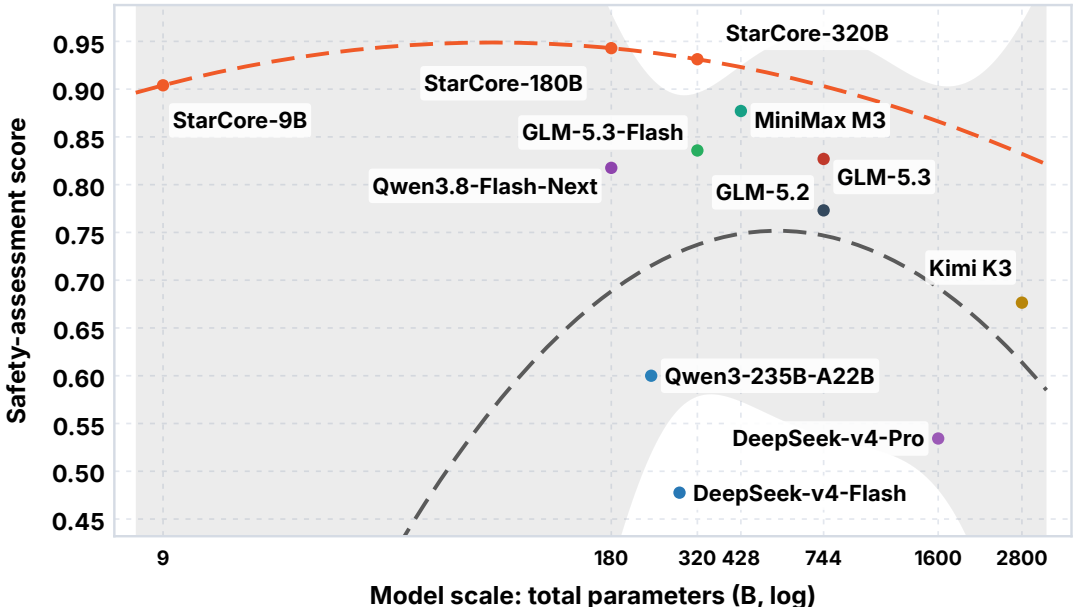


Figure 6: Safety-assessment score against total parameter count for twelve models (Colors distinguish models).

Fig. 6 displays a safety-assessment score against total parameter scale for twelve contemporary models. In particular, the StarCore [11] model is fine-tuned with state chains, a new format of training data. Therefore, we fit the performance of the StarCore model and the other LLMs separately. As displayed in the figure, the performance increases as the model size is small and achieves the optimal performance at the intermediate size. After that, the performance gradually decreases. This phenomenon validates our claim of condensing law of security information. Moreover, we observe that StarCore, which even uses a smaller model size, consistently outperforms existing LLMs on security tasks, suggesting that state chains provide a more effective format for security fine-tuning.

7 Conclusion

Scaling laws describe a world in which more parameters lead to greater capability, and for the most part they do. Security is the domain where that capability can work against these laws, because it causes the model to repair the very anomaly on which the security decision rests. Formalizing the trade-off as one between cognitive capability $K(W)$ and security-information retention $R(W)$ gives three theorems, which predict a U-shaped security generalization error in model size. We call the resulting statement the *condensing law of security information*: usable security information is unimodal in scale, so there is a finite $N^\star$ at which it peaks (condenses), and beyond which additional parameters reduce security performance rather than improve it. Experiments on the Pythia suite and real-world LLMs confirm our predictions: representations of normal and anomalous inputs converge as scale grows, and security-task performance peaks at an intermediate size.

References

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[3] Chaojun Xiao, Jie Cai, Weilin Zhao, Biyuan Lin, Guoyang Zeng, Jie Zhou, Zhi Zheng, Xu Han, Zhiyuan Liu, and Maosong Sun. Densing law of llms. *Nature Machine Intelligence*, 7(11):1823–1833, 2025.

[4] Šimon Morávek. HttpParamsDataset: values of HTTP request parameters, labelled benign or anomalous. `https://github.com/Morzeux/HttpParamsDataset`, 2016. Assembled for the author's diploma thesis from CSIC 2010, sqlmap, XSSYA, Vega Scanner and FuzzDB.

[5] Carmen Torrano Giménez, Alejandro Pérez Villegas, and Gonzalo Álvarez Marañón. HTTP DATASET CSIC 2010. Information Security Institute, Spanish National Research Council (CSIC), 2010.

[6] Bernardo Damele and Miroslav Stampar. sqlmap: automatic SQL injection and database takeover tool. `https://github.com/sqlmapproject/sqlmap`, 2025.

[7] FuzzDB Project. FuzzDB: attack patterns and primitives for black-box application fault injection. `https://github.com/fuzzdb-project/fuzzdb`, 2025.

[8] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[9] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2021.

[10] BigModelGym. LLM safety benchmark suite. `https://www.bigmodelgym.com/#overview`, 2026.

[11] StarCore. StarCore: The model that dreams the cyberspace world. `https://github.com/metaordo/StarCore`, 2026.

[12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

[13] Robert M. Fano. Transmission of information: A statistical theory of communications. *MIT Press and Wiley*, 1961.

A Proof of Theorem 1

By Assumption 1, $W \to P \to (X,A)$ is a Markov chain, so by the data-processing inequality [12]

$$I(W;X,A) \;\le\; I(W;P) \;=\; K(W). \tag{11}$$

By the chain rule,

$$I(W;X,A) \;=\; I(W;A) + I(W;X \mid A) \;\ge\; I(W;X \mid A). \tag{12}$$

Since $\widetilde{X} = \mathrm{Re}_W(A)$ is a function of (A,W), conditioning reduces entropy and Fano's inequality [13] gives

$$H(X \mid A,W) \;\le\; H(X \mid \widetilde{X}) \;\le\; \varepsilon \log|\Lambda| + 1. \tag{13}$$

From the relation between entropy and mutual information,

$$\begin{aligned} I(W;X \mid A) &= H(X \mid A) - H(X \mid A,W) \\ &\ge H(X \mid A) - \varepsilon \log|\Lambda| - 1. \end{aligned} \tag{14}$$

Combining the two inequalities,

$$K(W) \;\ge\; H(X \mid A) - \varepsilon \log|\Lambda| - 1, \tag{15}$$

and rearranging,

$$1 - \varepsilon \;\le\; \frac{K(W)}{\log|\Lambda|} - \frac{H(X \mid A) - 1}{\log|\Lambda|} + 1, \tag{16}$$

which is Eq. (3). ■

B Proof of Theorem 2

Since $Z = T_V(\widetilde{X})$, all security information in Z arrives through $\widetilde{X}$, and $S \leftarrow \Delta \rightarrow Z$ is a fork. By the data-processing inequality,

$$I(Z;S) \;\le\; I(\widetilde{X};S). \tag{17}$$

For $I(\widetilde{X};S)$,

$$I(\widetilde{X};S) \;\le\; I(\widetilde{X},X;S) \;=\; I(X;S) + I(\widetilde{X};S \mid X). \tag{18}$$

By Assumption 2, $I(X;S) = 0$, so

$$I(\widetilde{X};S) \;\le\; I(\widetilde{X};S \mid X) \;\le\; H(\widetilde{X} \mid X). \tag{19}$$

Since $\varepsilon = P(X \neq \widetilde{X})$, Fano's inequality applied with the roles of X and $\widetilde{X}$ exchanged gives

$$H(\widetilde{X} \mid X) \;\leq\; \varepsilon \log |\Lambda| + 1. \tag{20}$$

Therefore

$$R(W) \;=\; \frac{I(Z;S)}{I(A;S)} \;\leq\; \frac{\log |\Lambda|}{I(A;S)}\,\varepsilon + \frac{1}{I(A;S)}, \tag{21}$$

which is Eq. (4). ■

C Proof of Theorem 3

By Fano's inequality,

$$P(\widetilde{S} \neq S) \;\geq\; \frac{H(S \mid Z) - 1}{\log |\mathscr{S}|}. \tag{22}$$

Since $H(S \mid Z) = H(S) - I(S;Z)$ and $I(S;Z) = R(W)\, I(A;S)$,

$$P(\widetilde{S} \neq S) \;\geq\; \frac{H(S) - 1}{\log |\mathscr{S}|} - \frac{I(A;S)}{\log |\mathscr{S}|}\, R(W), \tag{23}$$

which is Eq. (5). ■