

Password Strength Analysis Through Social Network Data Exposure: A Combined Approach Relying on Data Reconstruction and Generative Models

Maurizio Atzori^{1,†}, Eleonora Calò^{2,*,†}, Loredana Caruccio^{2,†}, Stefano Cirillo^{2,†},
Giuseppe Polese^{2,†} and Giandomenico Solimando^{2,†}

¹Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale, 72, 09124, Cagliari (CA), Italy

²Department of Computer Science, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy

Abstract

Although passwords remain the primary defense against unauthorized access, users often tend to use passwords that are easy to remember. This behavior significantly increases security risks, also due to the fact that traditional password strength evaluation methods are often inadequate. In this discussion paper, we present SODA ADVANCE, a data reconstruction tool also designed to enhance evaluation processes related to the password strength. In particular, SODA ADVANCE integrates a specialized module aimed at evaluating password strength by leveraging publicly available data from multiple sources, including social media platforms. Moreover, we investigate the capabilities and risks associated with emerging Large Language Models (LLMs) in evaluating and generating passwords, respectively. Experimental assessments conducted with 100 real users demonstrate that LLMs can generate strong and personalized passwords possibly defined according to user profiles. Additionally, LLMs were shown to be effective in evaluating passwords, especially when they can take into account user profile data.

Keywords

Privacy-Preserving, Password-disclosure, Data wrapping, Data reconstruction, Social Network

1. Introduction

Traditional password strength assessments often fall short, as they focus on static syntax rules without considering the semantic context of user choices. Indeed, users generally choose passwords by using keywords easy to remember. However, since much personal information is shared on social networks, attackers can exploit these details to infer user passwords. Thus, through data reconstruction tools, it is possible to reconstruct information semantically related to a context close to users [1]. In this landscape, Large Language Models (LLMs) emerge as both a asset for evaluating password security and a potential threat in generating passwords. This discussion paper examines the privacy risks associated with sharing personal data online and explores the capabilities of LLMs in password evaluation and generation, as proposed in [2]. The latter presents SODA ADVANCE, an extension of the tool SODA [3], which includes a

SEBD 2025: 33rd Symposium on Advanced Database Systems, June 16-19, 2025, Ischia, Italy

*Corresponding author.

[†] These authors contributed equally.

✉ matzori@unica.it (M. Atzori); ecalo@unisa.it (E. Calò); lcaruccio@unisa.it (L. Caruccio); scirillo@unisa.it (S. Cirillo); gpolese@unisa.it (G. Polese); gsolimando@unisa.it (G. Solimando)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

new module for evaluating password strength based on information publicly available on social networks. This module exploits some approaches such as CUPP [4], LEET [5], COVERAGE [6], and FORCE [7], and introduces a new cumulative metric, namely Cumulative Password Strength (CPS). Furthermore, we present different pipelines, with aim of investigating capabilities and threats associated to the generation and evaluation of passwords by using different LLMs. The overall evaluation is driven by the following research questions (RQs):

- RQ1: Can we rely on LLMs to suggest complex and easy-to-remember passwords based on publicly available information on social networks?
- RQ2: Can LLMs represent a valid tool to support users in evaluating the strength of passwords based on personal information?
- RQ3: How does the public availability of personal information across multiple social networks impact the capabilities of LLMs to generate and evaluate password strength?
- RQ4: How effective is the prompt-based methodology for password generation and evaluation compared to state-of-the-art models?

2. Combining SODA ADVANCE and LLMs for Evaluating Passwords

In this section, we describe the SODA ADVANCE tool and the three proposed pipelines that combine the capabilities of LLMs¹ (e.g., Google Gemini, ChatGPT, Claude, Dolly, Falcon, and LLaMa) with those of SODA ADVANCE to address password generation and evaluation problems.

SODA ADVANCE Tool. The SODA ADVANCE tool evaluates password strength based on reconstructed personal data from social networks. The SODA ADVANCE pipeline (see Figure 1) starts with basic user information, i.e., name and photo as input (①). It then extracts public data from Facebook, LinkedIn and Instagram using web crawling and scraping techniques (②). The tool uses facial recognition to verify the user’s identity across all platforms (③). Finally, it merges the extracted information (④) and evaluates the strength of the provided password based on the reconstructed data (⑤). The evaluation module in SODA ADVANCE uses four methods (i.e., CUPP, LEET, COVERAGE and FORCE) and a new metric CPS that combines their results to provide a cumulative value in the range $[0, 1]$.

Generation and Evaluation Pipelines. The first pipeline is designed to investigate the capabilities of LLMs to generate strong passwords based on specific information provided by users. The process begins with the generation of passwords using LLMs, where each template creates a set of strong but memorable passwords based on user input. The generated passwords are then evaluated with the SODA ADVANCE module, which analyzes their strength. Consequently, each password is labeled as weak or strong according to the strength score.

The second pipeline is designed to investigate the effectiveness of LLMs in assessing the strength of passwords by also considering their semantics in relation to user data. The process begins by generating strong passwords using the best LLM of the Generation Pipeline.

¹www.deepmind.google, www.chat.openai.com, www.claude.ai, www.databricks.com, www.falconllm.tii.ae, and www.llama.meta.com

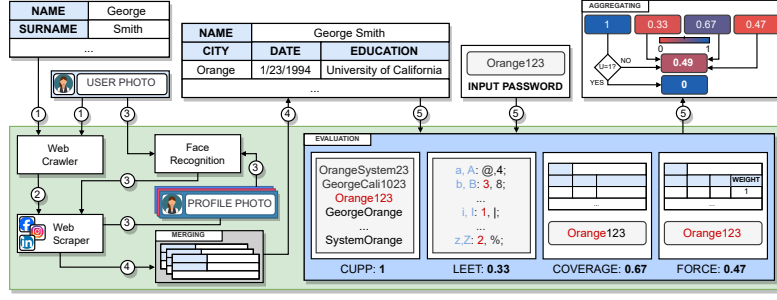


Figure 1: Overview of the modules underlying SODA ADVANCE.

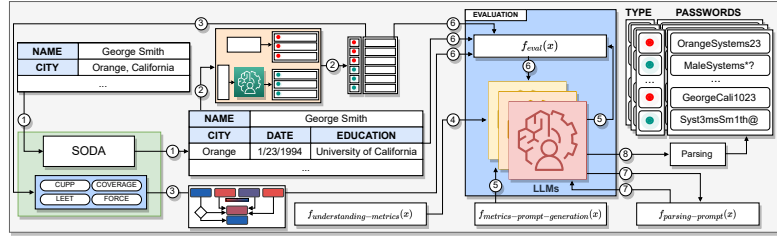


Figure 2: Overview of the data reconstruction and password evaluation pipeline.

Simultaneously, weak passwords are generated by using CUPP. Once passwords are created, a new prompt is generated to evaluate their strength. The evaluation involves submitting the user data along with the generated passwords to an LLM, which then assigns each password a numerical strength score. Finally, passwords are categorized as weak or strong according to obtained score. Details concerning the above-described pipelines can be found in [2].

Data Reconstruction and Password Evaluation Pipeline. The third pipeline combines the password strength evaluation of SODA ADVANCE with that of the LLMs, using new automated prompting functions for evaluating passwords. They directly consider within a prompt both data reconstructed from the social networks and the results achieved by SODA ADVANCE. As shown in Figure 2, starting from a small set of user information, we used SODA ADVANCE to reconstruct it using the publicly available information shared on social networks ①. Then, in ②, the reconstructed information is used to create a dataset containing both strong and weak passwords associated with the user. In ③, the set of user passwords is provided to SODA ADVANCE that is responsible for their first evaluation. Before proceeding with the evaluation step, in ④, new prompt containing the explanation of each metrics adopted by SODA ADVANCE is provided to LLMs. Moreover, for each of them, in ⑤, a new prompt considered both values resulting from SODA ADVANCE and the data reconstructed from the social networks, is automatically generated, which is then submitted to LM together with the passwords to be evaluated in ⑥, each prompt is filled with the user reconstructed data and the evaluation results from SODA ADVANCE, and it is submitted to an LLM together with the passwords to be evaluated. Finally, in ⑦ and ⑧, each password is associated a strength score to identify its category: strong or weak.

Prompt Engineering Approach for Password Strength problem. The process of generating passwords required the definition of an ad-hoc prompting function, namely $f_{\text{password-generation}}$, as shown in the following.

On the basis of the following personal information: [Name: George], [Surname: Smith], [City: Orange, California], [Date: 10/23/1994]. Could you generate a set of passwords that do not have to directly contain personal data, but must be easy for the user to memorize?

Instead, the process behind the password evaluation pipeline requires interacting with LLMs at several steps. Among these, we defined a new function $f_{\text{prompt-generation}}$ to ask each LLM to automatically generate prompts for password evaluation and a new function $f_{\text{parsing-prompt}}$ to ask each LLM to provide a strength score for each textual description. These prompts enabled us to automatically create a new prompting function, namely $f_{\text{evaluate-password}}$ for each LLM involved in our study. An example of the prompt automatically generated for ChatGPT follows:

User information: [Name: George], [Surname: Smith], [City: Orange, California], [Date: 10/23/1994], [Education: University of California]. For each line containing a password that I could use for a social network account, give me an answer for each of them and write whether the password can be considered secure or not, giving secure or not secure. Assess the password's strength using the information supplied by the user, considering factors like its length and ability to resist guessing techniques. Passwords: [OrangeSystems23], [MaleSystems*?], [GeorgeCali1023], [C@liforn1Sm1th49], [Syst3msSm1th@], [0r@nge@n3@]

In the third pipeline the interaction with LLMs to evaluate the password strength has required the use of some of the previous prompting functions, and the definition of new ones to explain the metrics (i.e., $f_{\text{understanding-metrics}}$) to LLMs and evaluate the password, by also considering the results of SODA ADVANCE. We manually defined two new prompting functions following the *Manual Template Engineering* strategy [8], and we automatically generated those to evaluate passwords for each LLMs, by means of $f_{\text{metrics-prompt-generation}}$ function. Starting from the generated function f_{eval} , we automatically generate a new specific prompt is provided to each LLM. The prompt generated by ChatGPT is shown below:

User information: [Name: George, Surname: Smith, City: Orange, California, Date: 10/23/1994]
 Passwords Evaluation Results:
 Password; Force; Leet; Coverage; CUPP; CPS
 OrangeSystems; 23; 57; 57; 0; 0.45
 MaleSystems*?; 27; 2; 71; 1; 1
 GeorgeCali1023; 63; 12; 76; 0; 0.50
 C@liforn1Sm1th49; 65; 0; 83; 0; 0.49
 Please assess the security of each password listed. Using the user information provided, analyze the password strength based on the following methods: Leet Coverage, Force, CUPP, and Cumulative Password Strength. Upon evaluation, please provide a response of Strong if the password is deemed sufficiently strong and effectively safeguards the user's information based on the provided data, or Weak if the password could potentially be compromised or guessed based on the available details.

The prompts generated by LLMs for evaluating password strength have showed similarities in their structures but have demonstrated differences in formatting and language style. In the following sections, we will show a case study involving real users that allows us to investigate the capabilities of SODA ADVANCE and LLMs to evaluate password strength.

3. Experimental Evaluation

The experiments in this study aim to evaluate how password strength can be affected by the information publicly available on social network platforms from both syntactical and semantic perspectives. To this end, we investigate the behavior of SODA ADVANCE and generative LLMs following the three different pipelines discussed in the previous section. We involved 100 users, each of whom filled out an information survey and an authorization form for profiling their social network using SODA ADVANCE. Among the questions submitted to users, we required their name, surname, and a photo. The collected data is used as starting points of the evaluation. Notice that, we obtained the explicit consent by users, in compliance with GDPR [9].

Technical Settings. SODA ADVANCE was implemented using Python version 3.10.2 on the server side and using web programming frameworks for graphical interfaces. Concerning LLMs, we adopt ChatGPT 3.5.5, Claude 2.1, LLaMa 2024.2.19.1, Falcon in its version at 40B, Google Gemini 1.0, and Dolly-v2-12b. Moreover, for the analysis of the characteristics of the generated passwords we used two different tools (i.e., *Passat* and *Node-password-analyzer*)². Furthermore, to make a comparative evaluation with SODA ADVANCE, we use the *Zxcvbn library* [10] in its version 4.4.2, the *CKL_PSM library* [11], and the *Semantic PCFG* [12] tool. The latter tool was trained on plain text passwords extracted from the *Evite*³ dataset. Finally, for generative password comparison, we use the *PassBERT model* [13].

RQ1. The characteristics of the generated passwords revealed that each LLM exhibits distinct patterns in the generation of strong passwords, with variations in syntactical complexity and the combination of letters/characters. Thus, to evaluate the strength of passwords we used the new metric CPS of SODA ADVANCE. In average, we obtained that Claude, Google Gemini, and ChatGPT outperform the other LLMs achieving a score of 0.82, 0.75, and 0.74, respectively. On the other hand, Dolly, LLaMa, and Falcon have generated more weak passwords, achieving a score of 0.65, 0.66, and 0.66, respectively. This is probably due to their tendency to generate repetitive or predictable passwords, using recurring and easily guessable patterns.

RQ2. Starting from the values provide by CPS, we consider a password as *strong* when its strength score is greater than or equal to 0.55, *weak* otherwise. Those, we are able to get a binary evaluation of passwords and compare the results achieved by LLMs with those achieved by methods proposed in the state-of-the-art. By considering the average value achieved by each LLM, Claude obtained the highest values for accuracy, precision, recall, and F1-score, i.e., 0.75, 0.76, 0.75, and 0.75, respectively. The high precision score indicates that it has a low rate of False Positive, meaning that it correctly identifies strong passwords with a high degree of confidence. To further investigate if the ensemble of different LLMs improves the values of metrics, we considered two different ensembles: *i*) including all the LLMs and *ii*) including the three LLMs with the highest scores; but both performed lower than Claude.

²www.github.com/HynekPetrak/passat, www.github.com/T-PWK/node-password-analyzer

³www.haveibeenpwned.com

RQ3. By combining social media data with the semantic capabilities of LLMs, password strength evaluations significantly improved with respect to scenario in which a few user data is provided to LLMs. Compared to the latter scenario, the inclusion of broader personal information led to better performance across most models. For instance, Falcon improved its precision from 0.48 to 0.77 and ChatGPT reached high scores in accuracy, precision, recall, and F1-score. Instead, Claude showed the best overall performance (i.e., accuracy equal to 0.77 and precision equal to 0.89). Ensemble models also benefited, likely due to the enhanced performance of individual LLMs. These improvements suggest that public social media data provides valuable context, allowing LLMs to make more accurate assessments. However, this also raises privacy concerns: as more personal data becomes accessible, users face increased risks. LLMs could be exploited by attackers to guess passwords based on publicly shared information. This highlights the importance of strong privacy settings, secure password practices, and the need for clear ethical and legal guidelines regarding the use of LLMs.

RQ4. To evaluate the capabilities of LLMs in both password generation and evaluation tasks, as well as the effectiveness of SODA ADVANCE in assessing password strength, we analyzed the medium-security passwords and compared the results with state-of-the-art tools.

Medium Password Strength evaluation. Starting from the initial dataset provided by 100 users, we generated a set of 30 passwords for each user using the prompt $f_{\text{password-generation}}$. The values of CPS obtained for medium-strength passwords generated by LLMs and evaluated through SODA ADVANCE range between 0.36 and 0.60. In particular, Claude, Google Gemini, and ChatGPT outperform all other LLMs achieving the highest number of medium passwords. Then, to assess the evaluation capabilities of LLMs and SODA ADVANCE, we asked each model to evaluate each password. By using the evaluation pipeline, the classification task involving multiple labels (i.e., *weak*, *medium*, *strong*), significantly reduced the performance of all LLMs with respect to the binary classification task (i.e., *weak* and *strong*). In particular, we have noticed that most of the passwords correctly evaluated were weak passwords, containing recurrent patterns and combinations of user data. Instead, LLMs were not able to discriminate passwords between strong and medium levels. Conversely, with the data reconstruction and password evaluation pipeline, the overall performance was higher, demonstrating that Claude outperformed all other LLMs. Our analysis shows that the initial evaluation provided by SODA ADVANCE effectively supported LLMs in distinguishing between weak, medium, and strong passwords.

Comparative evaluation with state-of-the-art tools. We performed a comparison with SODA ADVANCE and some of the most recent tools for password evaluation available in the state-of-the-art, i.e., Zxcvbn, CKL_PSM, and Semantic PCFG. In order to be able to compare the values obtained from the library and tools with those of the evaluation module of SODA ADVANCE, we uniform the ranges to fit the strength of the passwords in three categories, *weak*, *medium*, and *strong*. For the purposes of our evaluation, we extracted a random sample of 250 passwords, ranging in length from 8 to 25 characters. Figure 3 shows the results of SODA ADVANCE, CKL_PSM, Zxcvbn, and Semantic PCFG on the considered set of passwords. As we can see, most of the passwords have been classified as medium by all tools, and only a few of them as strong. SODA ADVANCE has demonstrated good capabilities of evaluation for the passwords containing these types of information, classifying them as weak. Moreover, SODA ADVANCE classified as

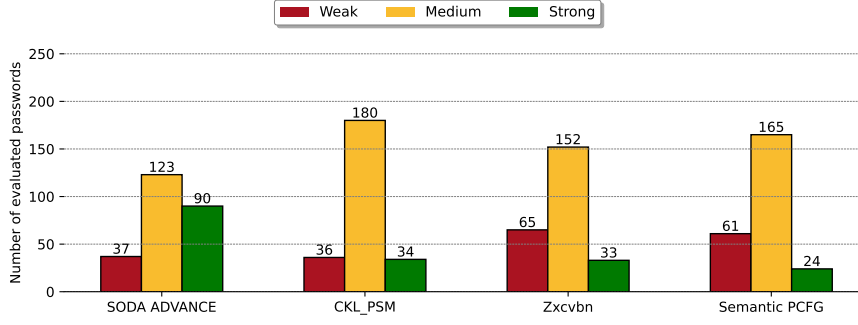


Figure 3: Result obtained from SODA ADVANCE, CKL_PSM, Zxcvbn, and Semantic PCFG on *Evite* dataset.

medium some passwords consisting of simple dictionary words not semantically linked to users. These types of passwords have been considered strong by the methods that evaluate these attempts, i.e., CKL_PSM, Zxcvbn, and Semantic PCFG, since they have a medium-complex syntax that requires a large number of attempts to crack. This is probably due to the metrics for the analysis of syntax included in the cps.

By summarizing, we have noticed that no model excels at evaluating password strength. As we expected, SODA ADVANCE demonstrated good evaluation capabilities for passwords that contain some user information but overestimates the complexity of passwords when they contain words not semantically linked to the user. On the other hand, tools that evaluate passwords based on crack attempts often underestimate the strength of passwords with complex syntax if they contain information related to the user. However, as also demonstrated for LLMs, considering the problem of evaluating password strength based on semantics with three levels of strength is extremely more difficult and the evaluations are less accurate.

Evaluating passwords with a state-of-the-art model. To further investigate the password-generation capabilities of LLMs, we evaluated the strength of the passwords with PassBERT [13], which is one of the most recent models in the literature for making focused attacks on passwords. PassBERT uses the fine-tuning paradigm for password-guessing attacks, with a pre-trained password model and different fine-tuning approaches. Among them, we considered Targeted Password Guessing (TPG) which aims to estimate the number of guesses of cracking the input password given a set of leaked passwords. For the purposes of our evaluation, we considered 100 users and their 250 strong passwords generated by LLMs. Moreover, we considered the weak passwords inferred by CUPP as leaked passwords. For each strong password, we evaluated its strength with the PassBERT model and the TPG approach. By considering 250 passwords for each user, we collected a total of 25,000 strong passwords. The results showed that among the strong passwords, only the passwords of a small set of users were inferred by PassBERT. Specifically, PassBERT was able to identify only 22 passwords out of the 25,000 evaluated, probably due to the complexity of the syntax of these passwords. In fact, although the passwords generated by LLMs are based on personal information about the user and therefore easy to remember, they are also syntactically complex and difficult to crack for models such as TPG. These results, together with those achieved from the previous evaluation, underscore the robustness of using LLMs for generating secure passwords semantically related

to the information of the users and highlight the limited effectiveness of an advanced targeted guessing model, i.e., PassBERT.

4. Conclusion and Future Directions

We have investigated the threats related to the definition of password when users publicly share their data on social network platforms. To this end, we have first proposed a new data reconstruction tool, namely SODA ADVANCE, capable of reconstructing public user data and evaluating a password according to them. Moreover, we have designed three different pipelines aiming to evaluate the performance of emerging LLMs, in the generation of strong passwords and the evaluation of their strength by a new ad-hoc prompting functions based on automatic and manual prompt engineering approaches. The experimental evaluations with real users have shown that Claude revealed good capabilities in generating strong passwords and evaluating password strength based on user data. Moreover, the combination of LLMs with the SODA ADVANCE tool has led to significant improvements in the password evaluation process with LLMs. To further investigate the effectiveness of LLMs and SODA ADVANCE in password generation and evaluation, we compared it with state-of-the-art approaches. The results highlight that LLMs do not perform well in the generation of medium-level passwords. Instead, the evaluation methods included in SODA ADVANCE performed better in this task. Finally, it has been shown that a very small percentage of strong passwords generated by LLMs succeed in being leaked by PassBERT's TPG model.

The methodologies and results obtained in this study open the research in several new directions. Future research could investigate in-depth the understanding and mitigation of threats, including exploring alternative approaches to password management and authentication in the context of widespread public data availability. In addition, further investigation could focus on enhancing the capabilities of the data reconstruction tool to extract a large set of public information from other Web platforms. Moreover, password strength assessment can be further explored using LLM by investigating the effectiveness of models trained specifically for this problem. Finally, emerging trends related to LLMs require further investigation for a better understanding of how these models treat personal information and whether they comply with European and global regulations.

Acknowledgments

This work was partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

- [1] S. Cirillo, D. Desiato, M. Scalera, G. Solimando, A visual privacy tool to help users in preserving social network data, in: Proceedings of the Workshops, Work in Progress Demos and Doctoral Consortium at the IS-EUD 2023 co-located with the 9th International Symposium on End-User Development (IS-EUD 2023), Cagliari, Italy, June 6-8, 2023, volume 3408 of *CEUR Workshop Proceedings*, 2023, pp. 1–8.
- [2] M. Atzori, E. Calò, L. Caruccio, S. Cirillo, G. Polese, G. Solimando, Evaluating password strength based on information spread on social networks: A combined approach relying on data reconstruction and generative models, *Online Social Networks and Media* 42 (2024) 100278.
- [3] F. Cerruto, S. Cirillo, D. Desiato, S. Gambardella, G. Polese, Social network data analysis to highlight privacy threats in sharing data, *Journal of Big Data* (2022).
- [4] Mebus, Common user password profiler, 2019. URL: <https://github.com/Mebus/cupp>, accessed 20 March 2019.
- [5] W. Li, J. Zeng, Leet usage and its effect on password security, *IEEE Transactions on Information Forensics and Security* 16 (2021) 2130–2143.
- [6] Y. Li, H. Wang, K. Sun, Personal information in passwords and its security implications, *IEEE Transactions on Information Forensics and Security* 12 (2017) 2320–2333.
- [7] X. Cui, C. Li, Y. Qin, Y. Ding, A password strength evaluation algorithm based on sensitive personal information, in: Proceedings of the IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) 2020, IEEE, 2020, pp. 1542–1545.
- [8] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, *ACM Computing Surveys* 55 (2023) 1–35.
- [9] P. europeo e del Consiglio, Regolamento (ue) 2016/679 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, 2016.
- [10] D. L. Wheeler, zxcvbn: Low-Budget password strength estimation, in: Proceedings of the 25th USENIX Security Symposium (USENIX Security 16), USENIX Association, Austin, TX, 2016, pp. 157–173.
- [11] M. Xu, C. Wang, J. Yu, J. Zhang, K. Zhang, W. Han, Chunk-level password guessing: Towards modeling refined password composition representations, in: Y. Kim, J. Kim, G. Vigna, E. Shi (Eds.), Proceedings of the CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021, ACM, 2021, pp. 5–20. doi:10.1145/3460120.3484743.
- [12] R. Veras, C. Collins, J. Thorpe, A large-scale analysis of the semantic password model and linguistic patterns in passwords, *ACM Transactions on Privacy and Security* 24 (2021) 1–21. doi:10.1145/3448608.
- [13] M. Xu, J. Yu, X. Zhang, C. Wang, S. Zhang, H. Wu, W. Han, Improving real-world password guessing attacks via bi-directional transformers, in: Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), USENIX Association, Anaheim, CA, 2023, pp. 1001–1018.