TÜRKİYE BİLİŞİM DERNEĞİ 1971
IEEE Türkiye Section
Ankara / Türkiye
5-6 February 2026
5th International INFORMATICS and SOFTWARE ENGINEERING CONFERENCE
Large Language Models in Software Engineering
https://iisec.tbdkakademi.org.tr/2026/
TurkiyeBilisimDernegi
bilisim2026

How Prompt Design Affects LLM-Based Password Strength Evaluation: A Comparison Against zxcvbn

Tuğberk Kocatekin

Istanbul Arel University

Question

- Can small, deployable LLMs reliably evaluate passwords strength?

Large Language Models

- A neural network trained on massive datasets to guess the next word.
  - NLP tasks, agents, etc.
- Local, deployable LLMs:
  - Reliable *(no down-time)*
  - Cost efficient *(no cost per token)*
  - Private (*local*)
  - Ollama

Password Strength Meter

Aspect	zxcvbn	Rule-based	Entropy-based
Core idea	Pattern matching + attack cost estimation	Predefined composition rules	Mathematical randomness estimation
Evaluation method	Detects dictionary words, names, dates, patterns, repeats, leetspeak	Checks length, character classes, forbidden patterns	Computes entropy from character set size and length
Output	Score (0–4), crack time estimates, feedback	Pass/fail or score based on rules	Entropy value (bits)
Context awareness	High (understands human password habits)	Low (syntactic only)	None (purely statistical)

Passwords

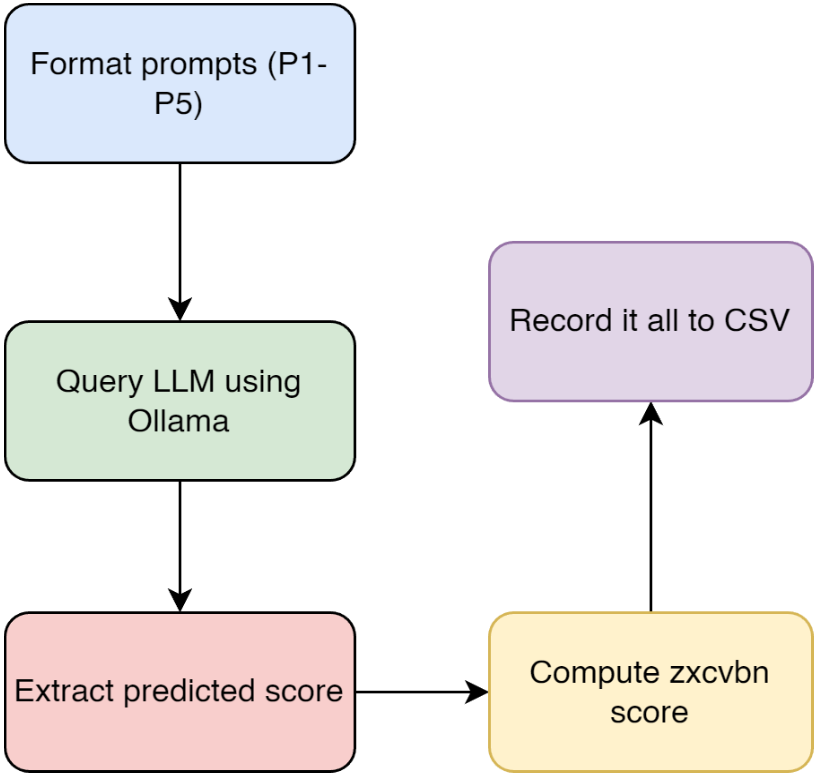
Score	Examples
0 – Very Weak	123456, qwerty, letmein
1 - Weak	love88, ros3bud99, Tianya
2 - Medium	l337speak, a6a4Aa8a, MorningWind7
3 - Strong	my_r3s3arc, FutureTech2, do you know
4 – Very strong	easytofindhard, canbeshortbutgood, correct-horse-battery-staple

Prompts

- **5 Prompts**
  - P1: Structured JSON
  - P2: Minimal
  - P3: Chain-of-thought
  - P4: Expert persona
  - P5: zxcvbn emulation

Method

- **Dataset**
  - 100 passwords (synthetic)
  - 20 per class (0-4)
- **Llama-3 8B via Ollama**
- **5 Prompts**
  - (P1) Structured JSON
  - (P2) Minimal
  - (P3) Chain-of-thought
  - (P4) Expert Persona
  - (P5) zxcvbn-emulation



Format prompts (P1-P5)

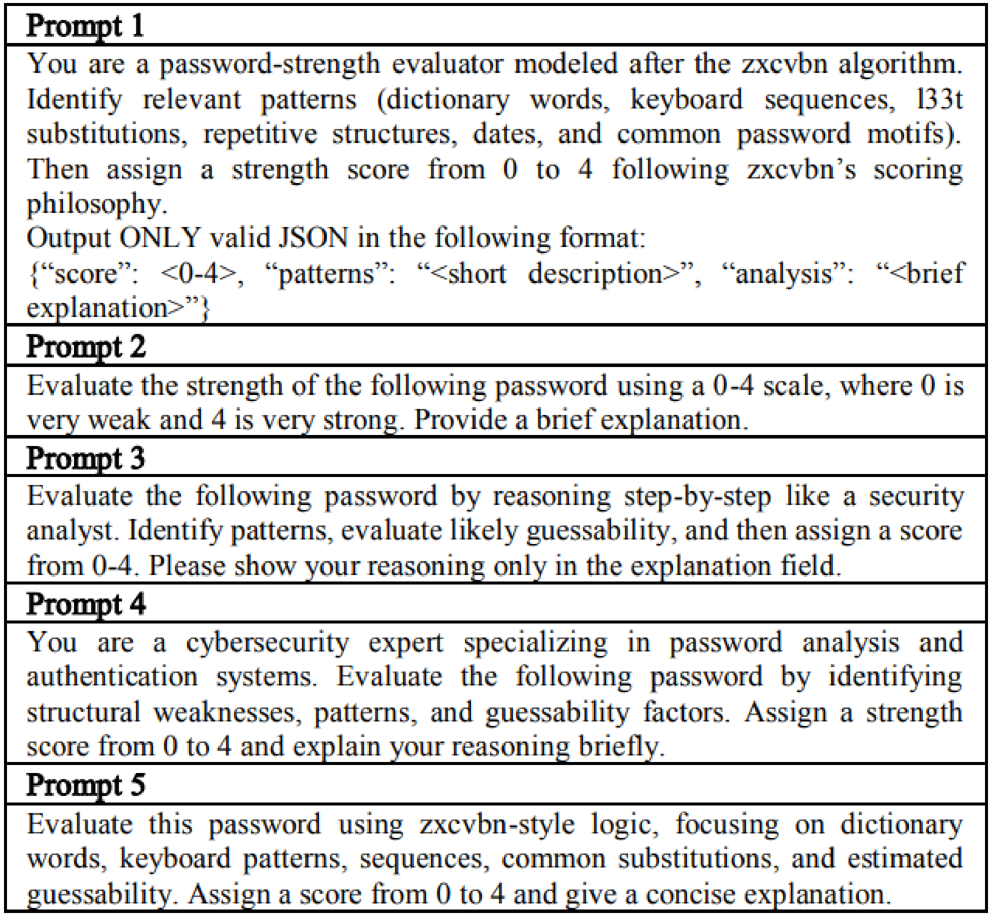
Query LLM using Ollama

Extract predicted score

Compute zxcvbn score

Record it all to CSV

Prompts



Prompt 1
You are a password-strength evaluator modeled after the zxcvbn algorithm. Identify relevant patterns (dictionary words, keyboard sequences, l33t substitutions, repetitive structures, dates, and common password motifs). Then assign a strength score from 0 to 4 following zxcvbn's scoring philosophy. Output ONLY valid JSON in the following format: {"score": <0-4>, "patterns": "<short description>", "analysis": "<brief explanation>"}
Prompt 2
Evaluate the strength of the following password using a 0-4 scale, where 0 is very weak and 4 is very strong. Provide a brief explanation.
Prompt 3
Evaluate the following password by reasoning step-by-step like a security analyst. Identify patterns, evaluate likely guessability, and then assign a score from 0-4. Please show your reasoning only in the explanation field.
Prompt 4
You are a cybersecurity expert specializing in password analysis and authentication systems. Evaluate the following password by identifying structural weaknesses, patterns, and guessability factors. Assign a strength score from 0 to 4 and explain your reasoning briefly.
Prompt 5
Evaluate this password using zxcvbn-style logic, focusing on dictionary words, keyboard patterns, sequences, common substitutions, and estimated guessability. Assign a score from 0 to 4 and give a concise explanation.

Evaluation Metrics

- Exact Agreement
- Soft Agreement
  - Catches near-misses
- Hallucination
  - Invalid outputs (> 4)

Results (Exact)

Prompts	Exact Agreement (%)				
	Score 0	*Score 1*	*Score 2*	*Score 3*	*Score 4*
P1	**6,7**	**25**	**80**	**5**	**0**
P2	40	35	50	15	0
P3	**0**	**35**	**5**	**10**	**5**
P4	6,7	45	10	10	0
P5	20	30	25	10	5

Results (Soft)

Prompts	Soft Agreement (%)				
	Score 0	*Score 1*	*Score 2*	*Score 3*	*Score 4*
P1	**6,7**	**100**	**100**	**80**	**45**
P2	40	90	95	80	70
P3	**0**	**60**	**75**	**20**	**30**
P4	6,7	75	70	72	0
P5	20	60	60	90	40

Results (Hallucination)

Prompts	Hallucination (%)				
	Score 0	*Score 1*	*Score 2*	*Score 3*	*Score 4*
P1	0	0	0	0	0
P2	0	0	0	0	0
P3	**20**	**35**	**15**	**5**	**0**
P4	20	10	15	5	0
P5	0	10	20	0	0

Discussion

- P1 and P2 are the **best** players.
  - Structured JSON & Minimal
  - No hallucination
  - Consistent results
- P3 (Chain-of-thought)
  - Worst
  - Unreliable (hallucination)
- P4 (Expert Persona)
  - Poor performance
  - Some hallucination

Limitations & Future Work

- **Single, small model**
  - Use larger models
- **Zxcvbn as reference**
  - **Not** ground truth.
- **Small, curated password dataset**
  - Use larger dataset
- **Decoding parameters are unexplored**
  - Try different temperatures
- **Prompts can be stricter**

Conclusion

- Small and deployable LMs are not drop-in replacements for deterministic password strength meters
    - They exhibit inconsistent scoring
- Chain-of-thought is unsafe due to hallucination risk
- Structured prompts improve reliability
    - Not as consistent as zxcvbn
    - No hallucinations