Validating Large Language Models with ReLM

Michael Kuchnik, Virginia Smith, George Amvrosiadis

Carnegie Mellon University

In Case You Don't Read the News

The Register

Machine learning models leak personal info if training data is compromised

Attackers can insert hidden samples to steal secrets

VICE Video TV News Tech Rec Room Food World News

Facebook's New AI System Has a 'High Propensity' for Racism and Bias

IBT

IBM's Watson Gets A 'Swear Filter' After Learning The Urban Dictionary

The Register

Microsoft's AI Bing also factually wrong, fabricated text during launch demo

Redmond's hype box and Google's Bard just as bad as each other

Email: A Curated Autocomplete

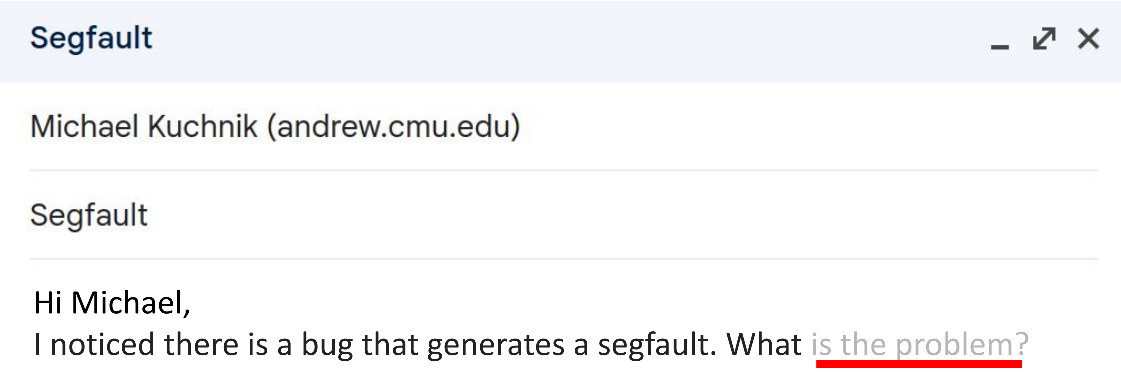
Segfault

Michael Kuchnik (andrew.cmu.edu)

Segfault

Hi Michael,
I noticed there is a bug that generates a segfault. What

Email: A Curated Autocomplete



Email: A Curated Autocomplete (GPT-2XL @ top_k=40)

Segfault

Michael Kuchnik (andrew.cmu.edu)

Segfault

Hi Michael,
I noticed there is a bug that generates a segfault. What is the problem?

Segfault

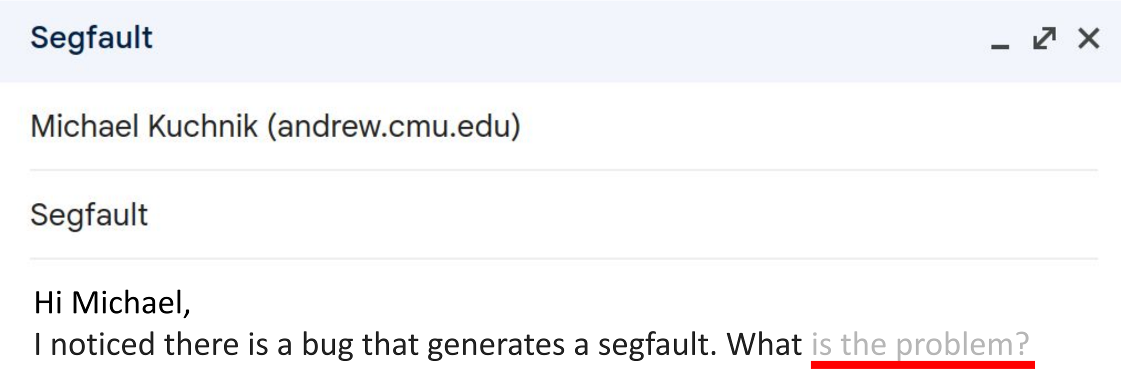
Michael Kuchnik (andrew.cmu.edu)

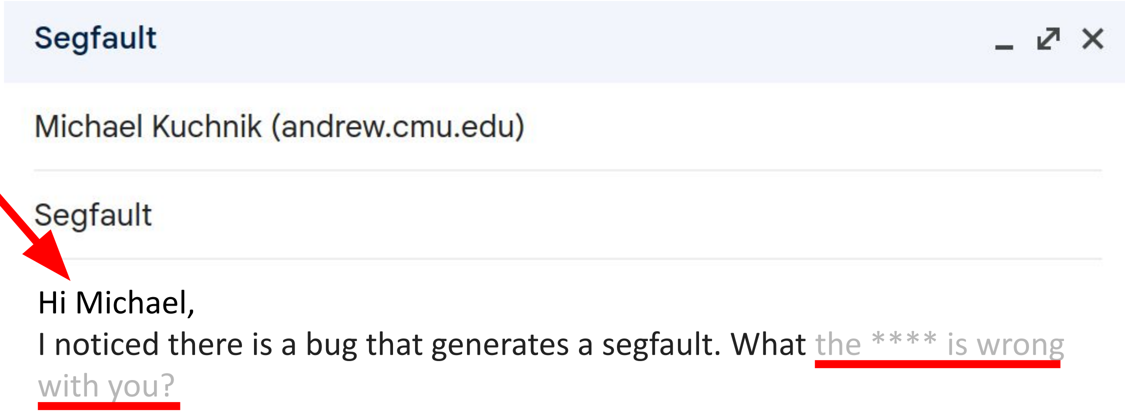
Segfault

Hi Michael,
I noticed there is a bug that generates a segfault. What the **** is wrong with you?

Email: A Curated Autocomplete (GPT-2XL @ top_k=40)

Biased outcome:
5x more likely with "Brittney"





Talk Synopsis: ReLM Helps Keep AI Models in Check

- Large Language Models (LLMs) can generate both good and bad content
  - Train once: $5+ million or more to train [1]
  - **Challenge:** Find + post-process "bugs"
- How to test model for errors?
  - **Current practice:** Sample random sentences or test next-token distribution
  - Hard to guarantee coverage is sufficient
- We built a tool, ReLM, to find bad content in LLMs
  - General purpose ("regex") queries against models
  - **Insight:** We can constrain the probability model itself

[1] https://news.mit.edu/2020/shrinking-deep-learning-carbon-footprint-0807

How to Test LLM Knowledge of George Washington?

George Washington was born on _____

George Washington

190 languages

Article Talk Read View source View history Tools

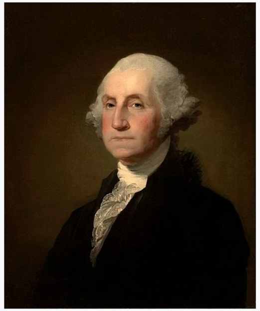
From Wikipedia, the free encyclopedia

"General Washington" redirects here. For other uses, see General Washington (disambiguation) and George Washington (disambiguation).

George Washington (February 22, 1732[b] – December 14, 1799) was an American military officer, statesman, and Founding Father who served as the first president of the United States from 1789 to 1797. Appointed by the Continental Congress as commander of the Continental Army, Washington led Patriot forces to victory in the American Revolutionary War and served as president of the Constitutional Convention of 1787, which created and ratified the Constitution of the United States and the American federal government. Washington has been called the "Father of his Country" for his manifold leadership in the nation's founding.[10]

Washington's first public office, from 1749 to 1750, was as surveyor of Culpeper County, Virginia. He subsequently received his first military training and was assigned command of the Virginia Regiment during the French and Indian War. He was later elected to the Virginia House of Burgesses and was named a delegate to the Continental Congress, where he was appointed Commanding General of the Continental Army and led American forces allied with France to a decisive victory over the British at the siege of Yorktown in

George Washington



Portrait by Gilbert Stuart, 1796

https://en.wikipedia.org/wiki/George_Washington

Designing a Test: Multiple Choice or Free Response?

George Washington was born on _____

A) February 23, 1973
B) March 30, 1973
C) February 1, 1873
D) February 22, 1732

Designing a Test: Multiple Choice or Free Response?

George Washington was born on _____

A) February 23, 1973
B) March 30, 1973
C) February 1, 1873
D) February 22, 1732

Designing a Test: Multiple Choice or Free Response?

George Washington was born on _____

A) ~~February 23, 1973~~
B) ~~March 30, 1973~~
C) ~~February 1, 1873~~
D) ~~**February 22, 1732**~~

A) January 1, 3000
B) January 1, 4000
C) January 1, 5000
D) **February 22, 1732**

Impossible: Hasn't happened yet

Designing a Test: Multiple Choice or Free Response?

George Washington was born on _____

A) ~~February 23, 1973~~
B) ~~March 30, 1973~~
C) ~~February 1, 1873~~
D) ~~February 22, 1732~~

A) ~~January 1, 3000~~
B) ~~January 1, 4000~~
C) ~~January 1, 5000~~
D) ~~February 22, 1732~~

Impossible: Hasn't happened yet

this day in 1732

Designing a Test: Multiple Choice or Free Response?

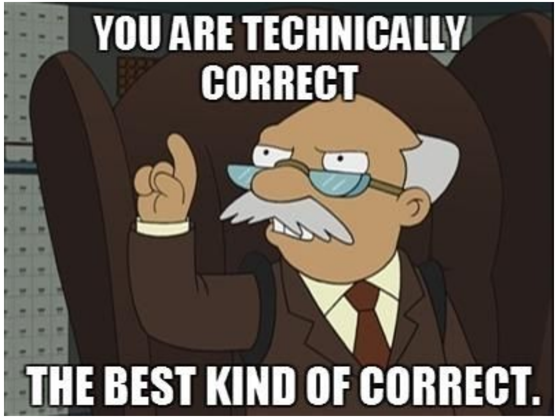
George Washington was born on _____

A) ~~February 23, 1973~~
B) ~~March 30, 1973~~
C) ~~February 1, 1873~~
D) ~~February 22, 1732~~

A) ~~January 1, 3000~~
B) ~~January 1, 4000~~
C) ~~January 1, 5000~~
D) ~~February 22, 1732~~

Impossible: Hasn't happened yet

~~this day in 1732~~
a farm



Takeaway: Limited Test Resolution or Uncontrollability

Avoiding Test Error with Constrained Decoding (ReLM)

George Washington was born on <date>

*<date> of the form <month> <day>, <year>

Avoiding Test Error with Constrained Decoding (ReLM)

George Washington was born on <date>
July 4, 1732

*<date> of the form <month> <day>, <year>

ReLM: <u>R</u>egular <u>E</u>xpressions for <u>L</u>anguage <u>M</u>odels

Goal of ReLM: Reduce Validation to Regex

- Designed as a single system for LLM validation
  - Targets efficient query execution, not query design
- Covering a broad range of tasks
  - **Memorization:** extraction of potentially private content
  - **Bias:** characterization of distribution
  - **Toxicity:** extraction of offensive content
  - **Language and Factual Understanding**: typical NLP benchmarks
- Focusing on efficiency of query execution
  - **Metrics:** Throughput, data efficiency, expressiveness
  - **SOTA:** hand-rolled unit-tests for LLM property

Regular Expressions: Regexs

- A regex represents a string pattern
- **Operations:**
  - Literal string (“I am a literal”)
  - OR (|)
  - 0+ repetitions (*)
- **Example:** The answer is ____
  - **Multiple Choice:** The answer is ((cat)|(dog))
  - **Free Response:** The answer is [a-z]*

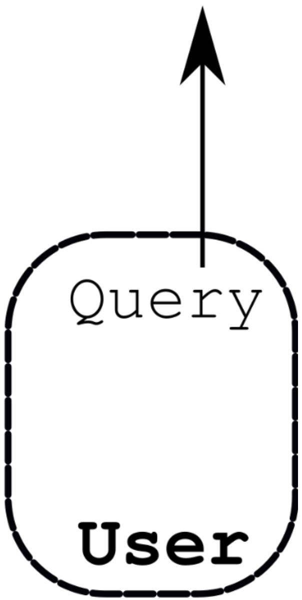
API: Constrained Decoding with ReLM

```
results = relm.search(
        model,
        query=("George Washington was born on "
                "(January|February|...) [0-9]{1,2}, [0-9]{4}"),
         # More Options
)
for x in results:
      print(x)  # George Washington was born on July 4, 1732
```

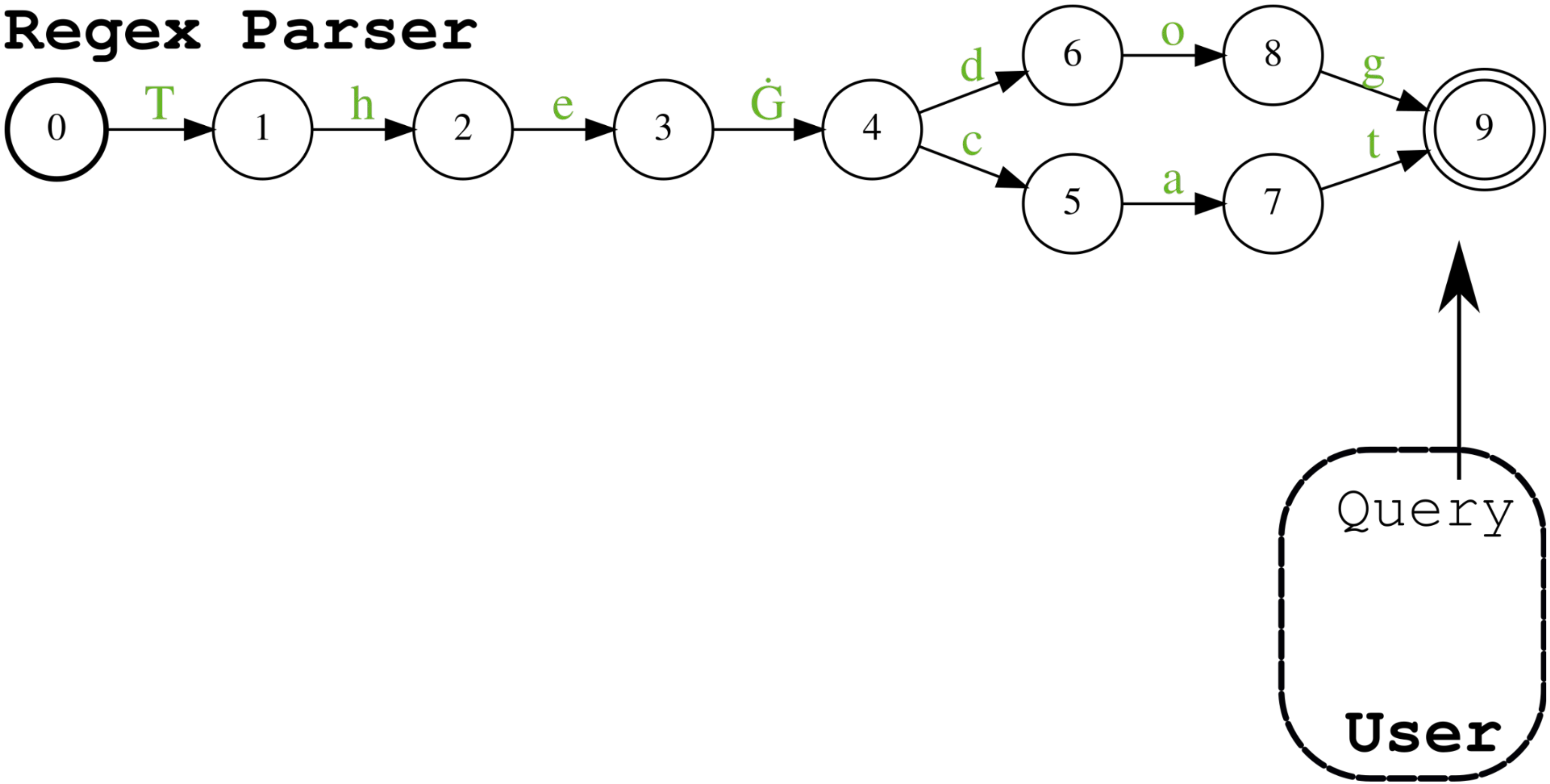
Designed by ML researcher

Argmax vs. Random
Exact vs. Fuzzy

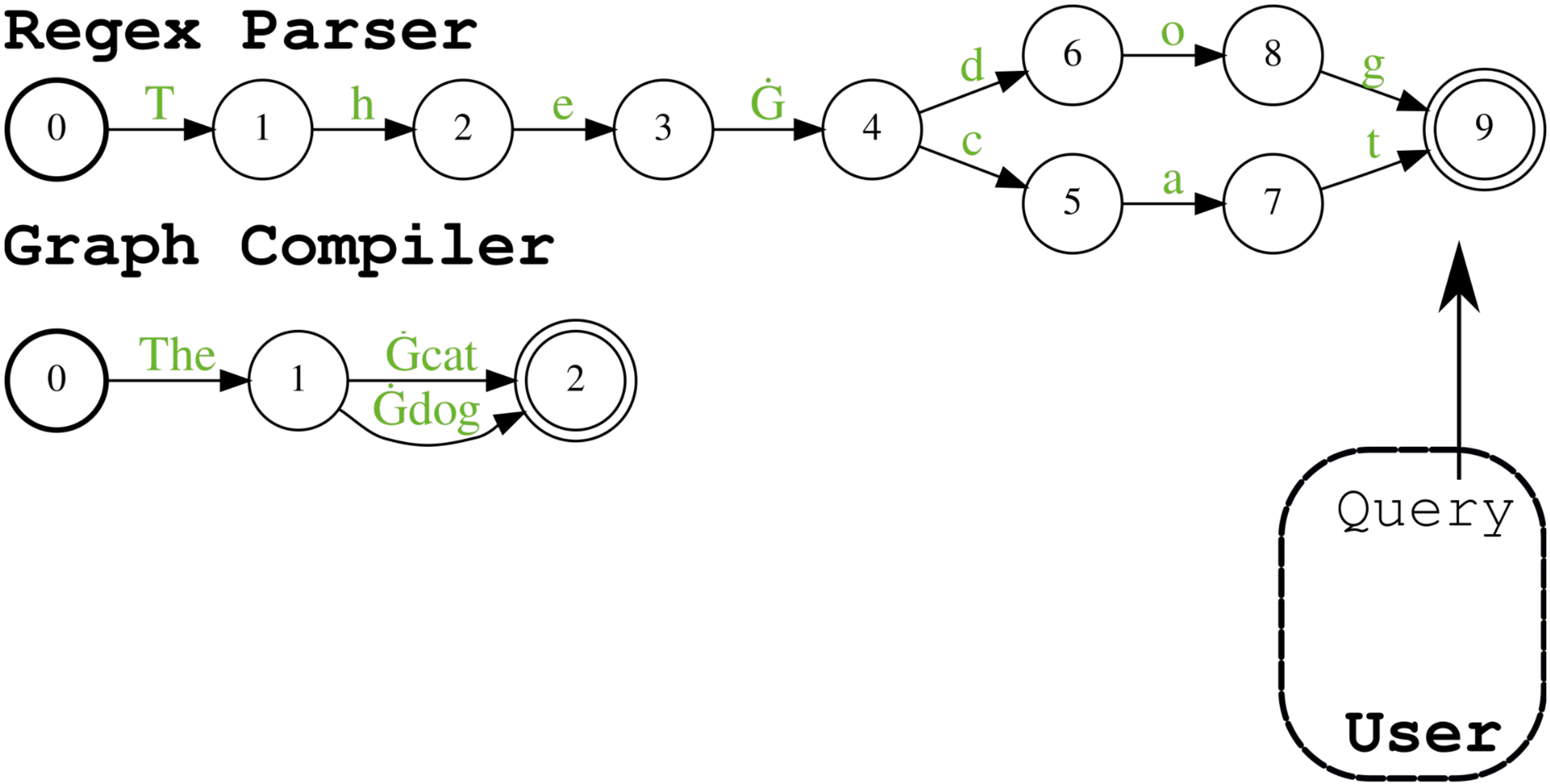
The ReLM WorkFlow for "The cat" or "The dog"



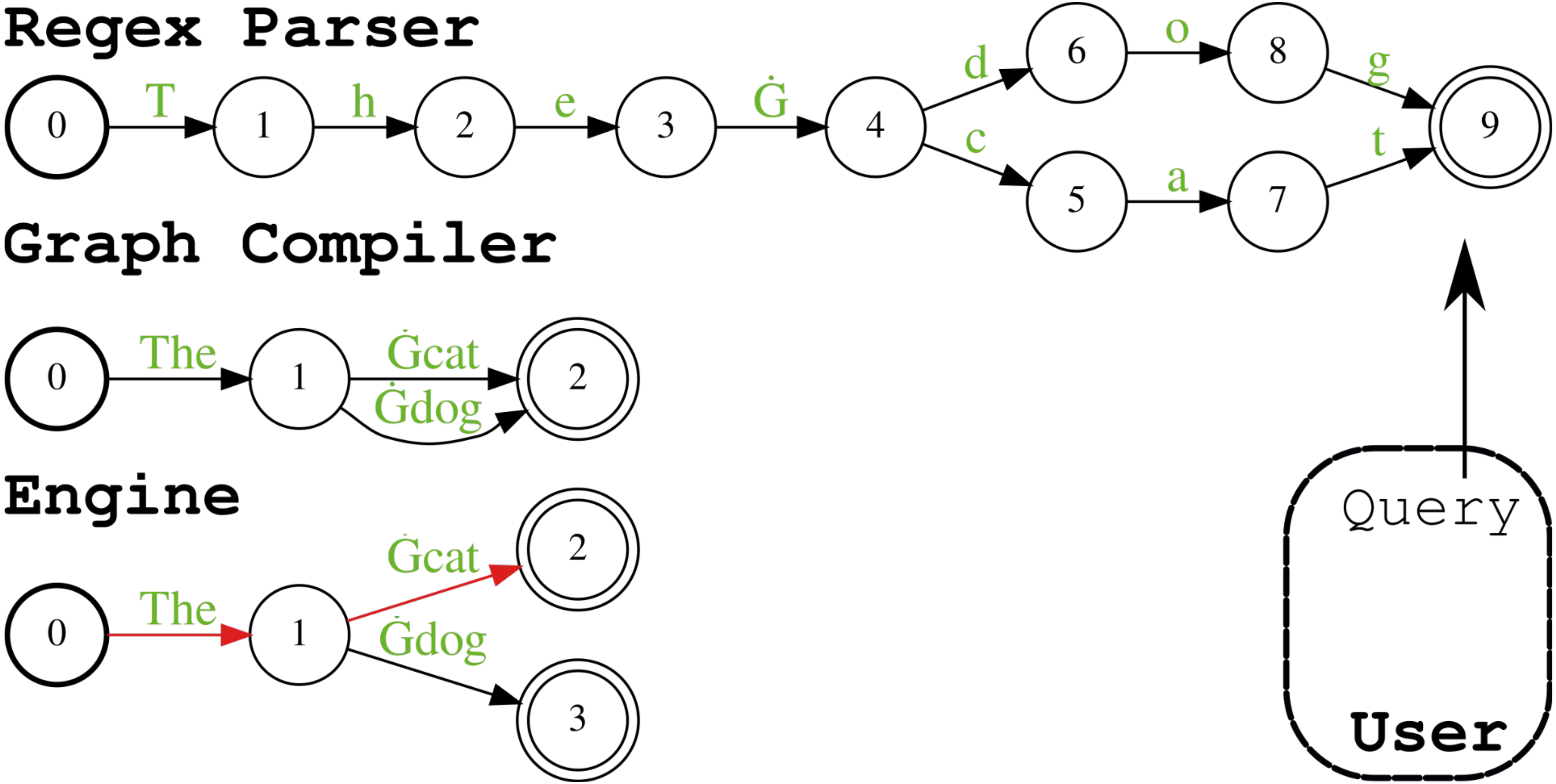
The ReLM WorkFlow for “The cat” or “The dog”



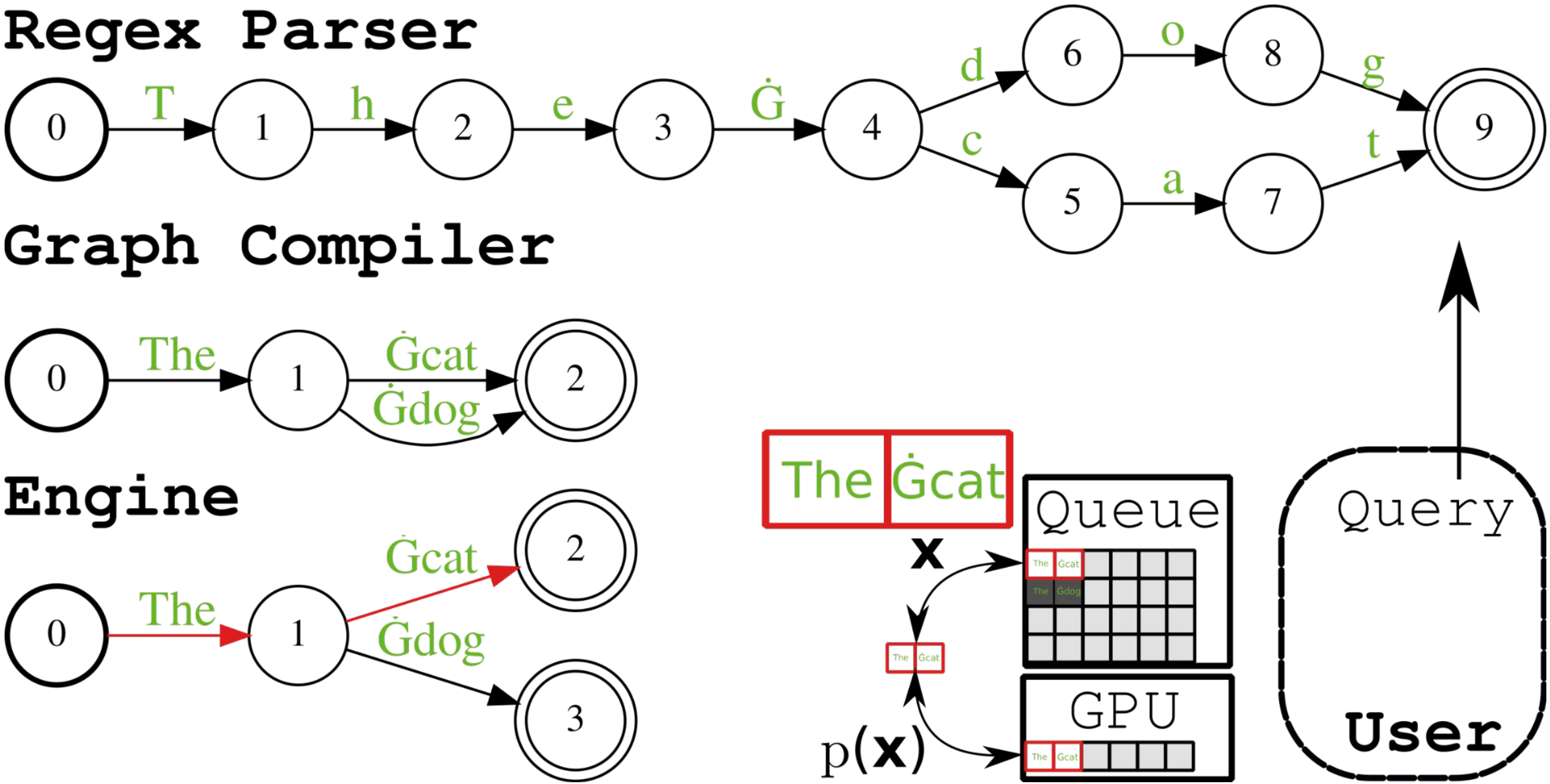
The ReLM WorkFlow for “The cat” or “The dog”



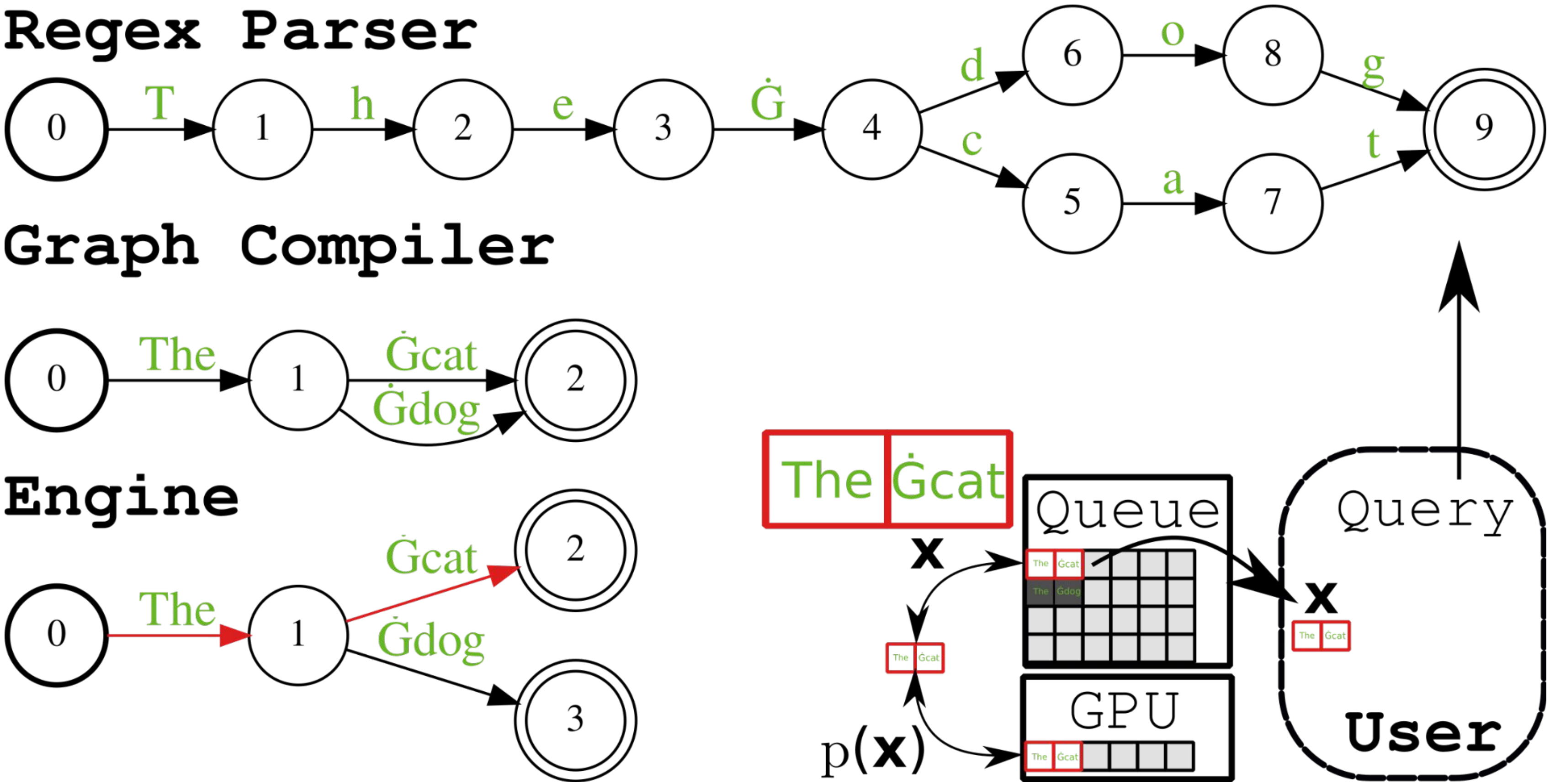
The ReLM WorkFlow for “The cat” or “The dog”



The ReLM WorkFlow for “The cat” or “The dog”

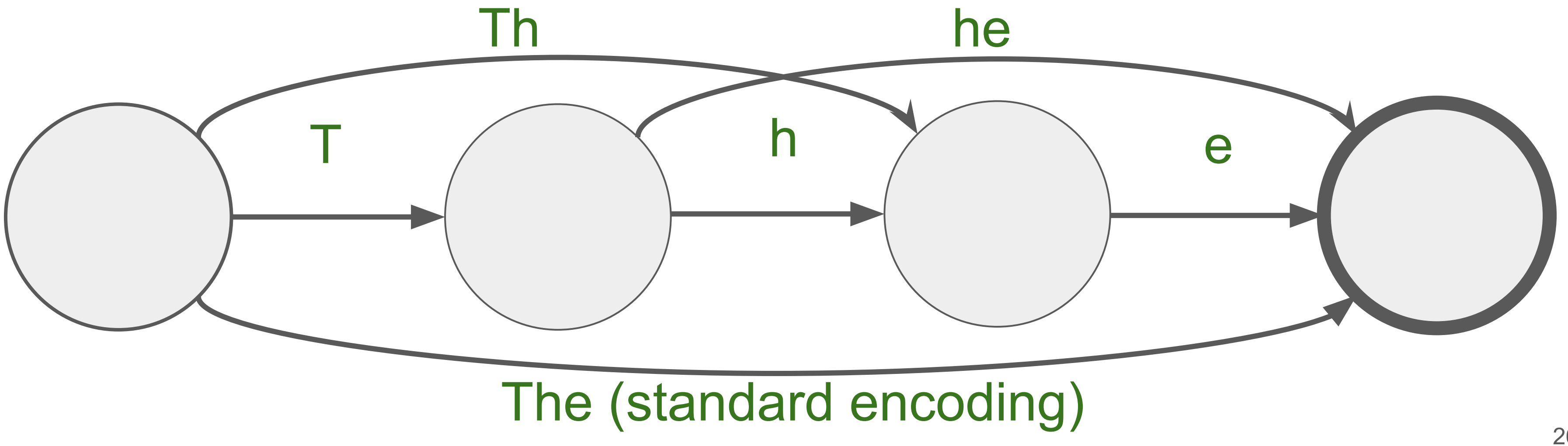


The ReLM WorkFlow for “The cat” or “The dog”



ReLM Compiles to Language of LLMs

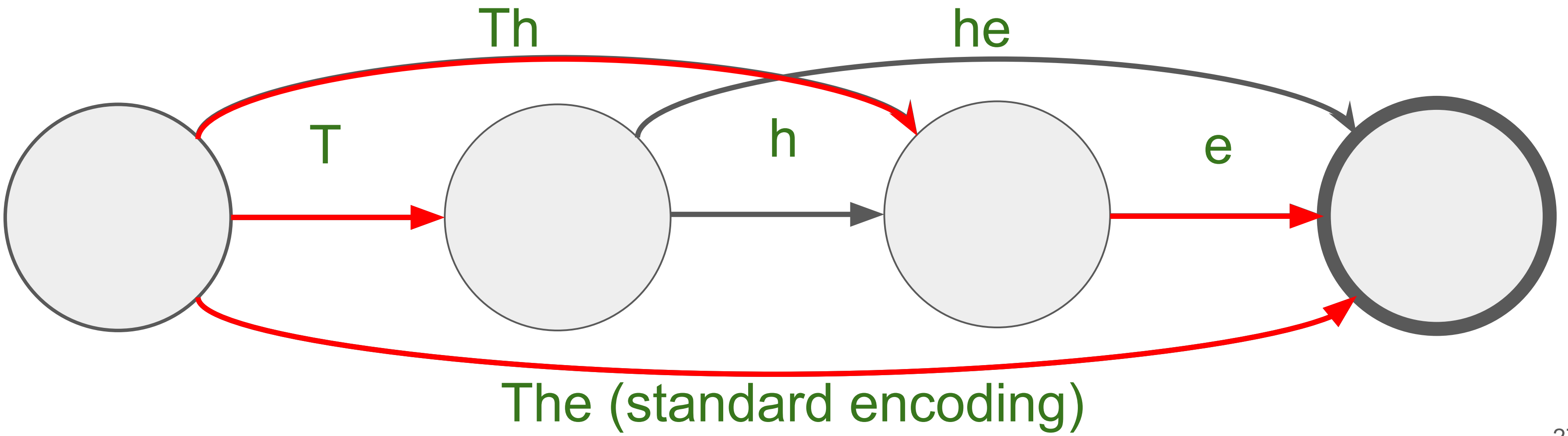
- A LLM operates over tokens (i.e., integers)
  - Tokens represent whole words, subwords, or characters
  - **Observation:** More than one encoding for a word



Extracting Matches from LLM with ReLM

Top-k: Filter all but top results at edge (k=2 shown)

Traversal: random or shortest paths in regex



Evaluation

Q1: Can Regex Express Validation Tasks?

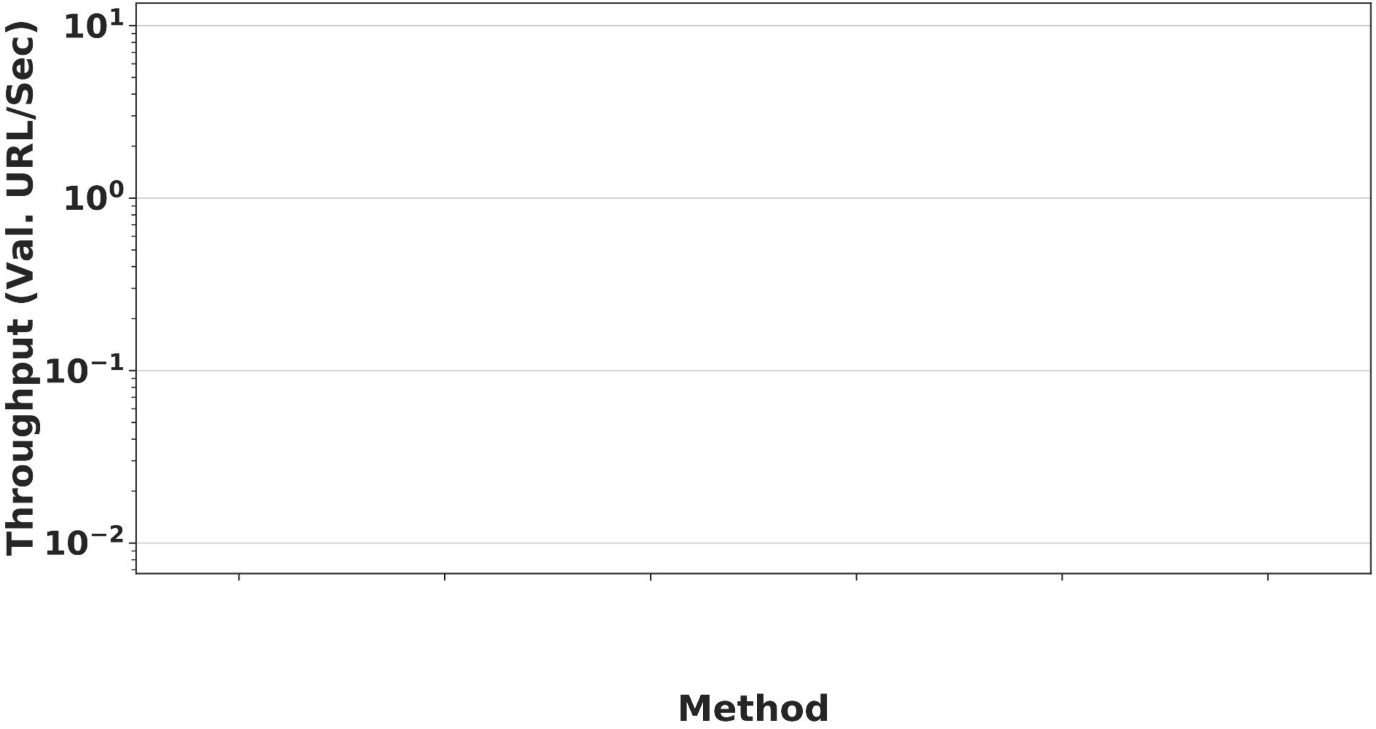
Q2: Can ReLM Outperform Baseline Implementations?

Setup: GPT2-XL@top_k=40, 1 GPU (Nvidia 3080)

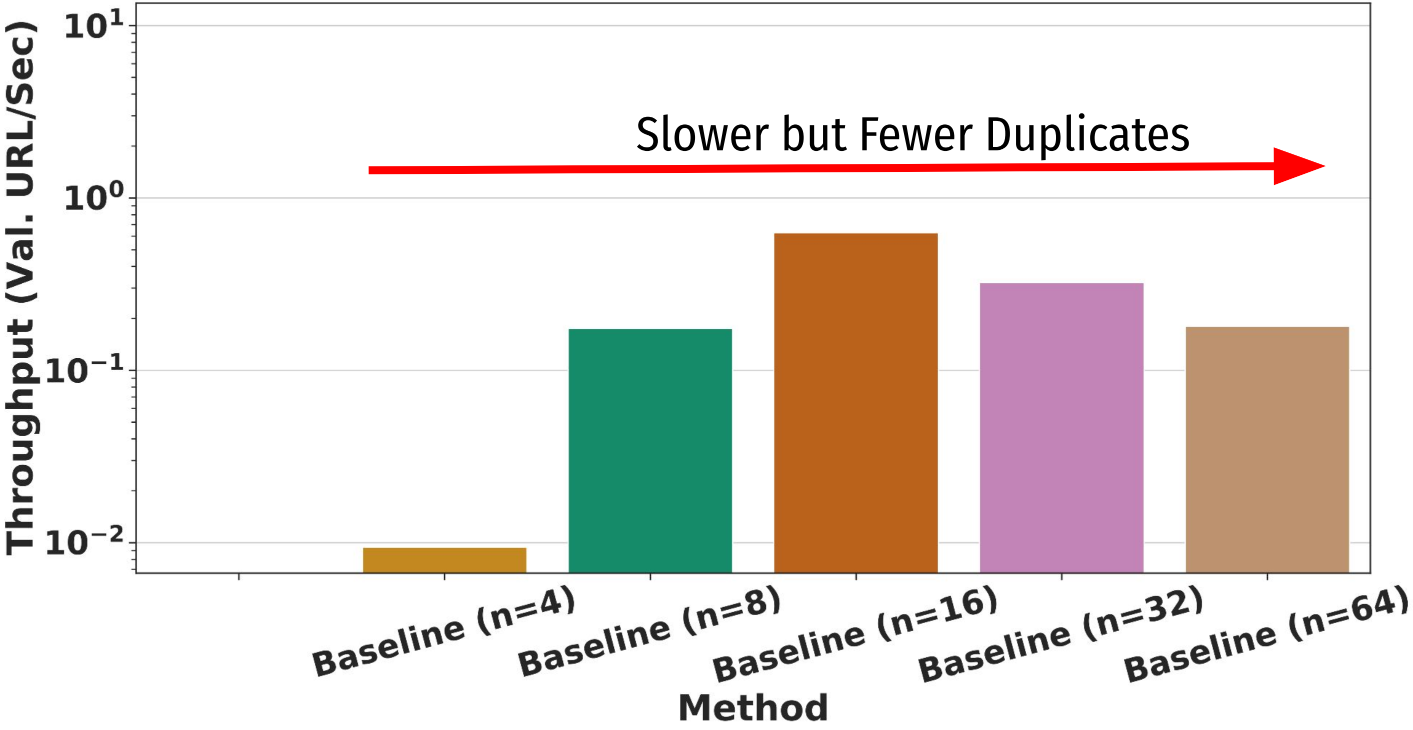
Evaluation: URL Memorization

- **Task:** Extracting valid URL
- **Pattern:** https://www.**<mydomain.topleveldomain/content>**
- **Accuracy Metric:** If URL resolves
- **Baseline:** Randomly sampling sequences of length up to **n**
- **ReLM:** Matching strings in most-likely order

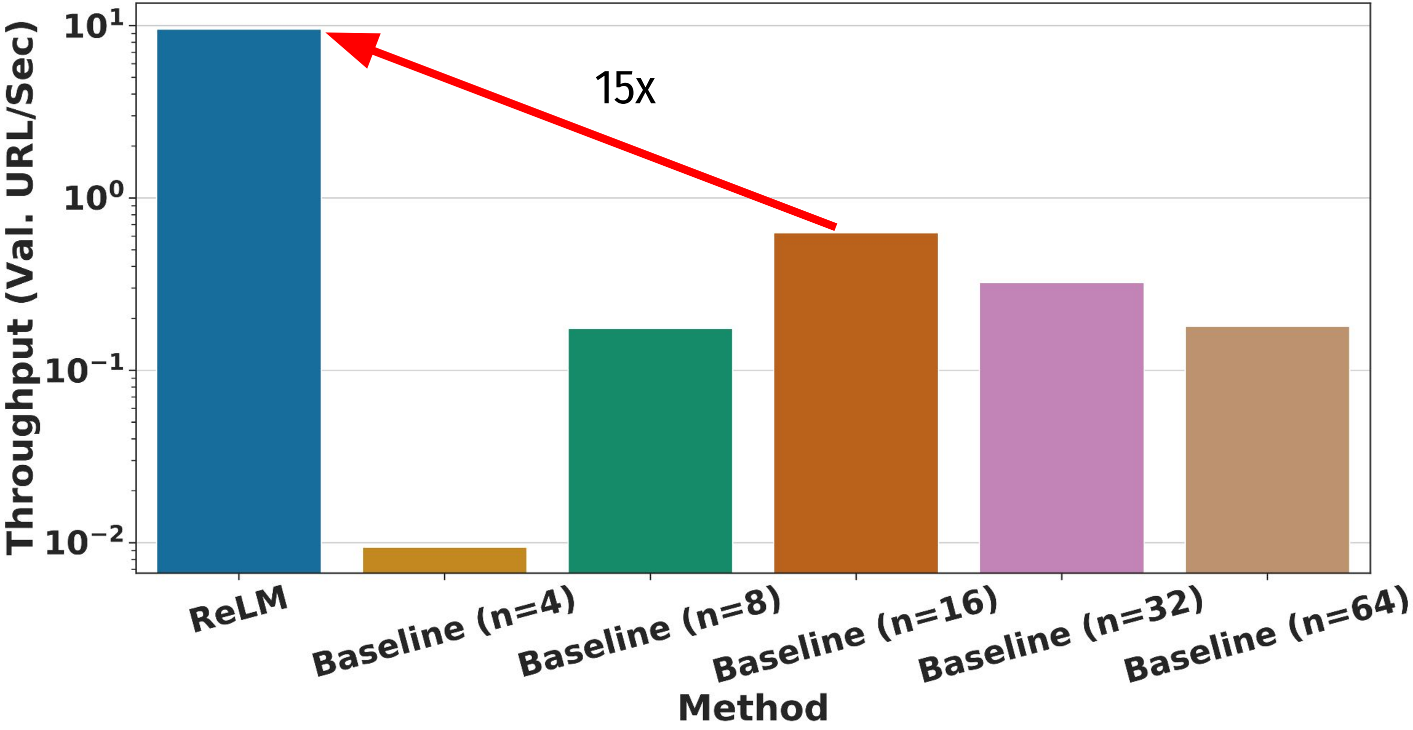
ReLM: 15x System Efficiency For Extractions



ReLM: 15x System Efficiency For Extractions



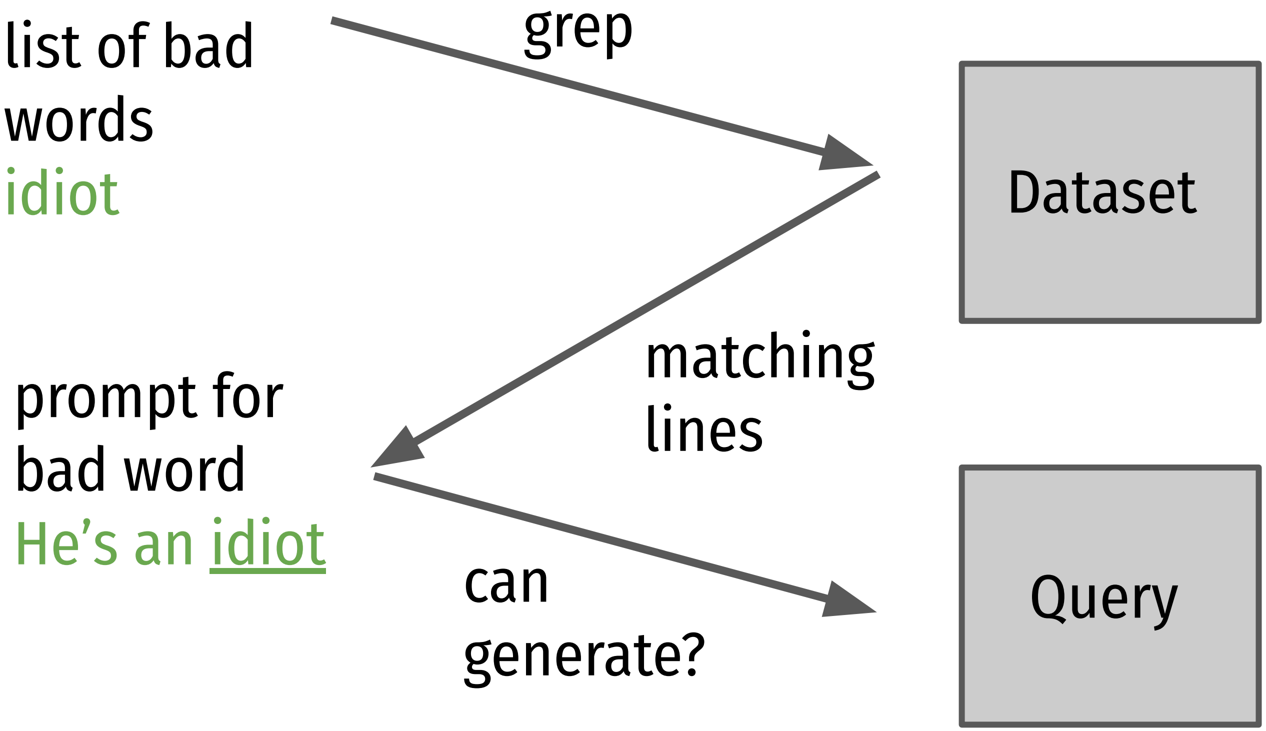
ReLM: 15x System Efficiency For Extractions



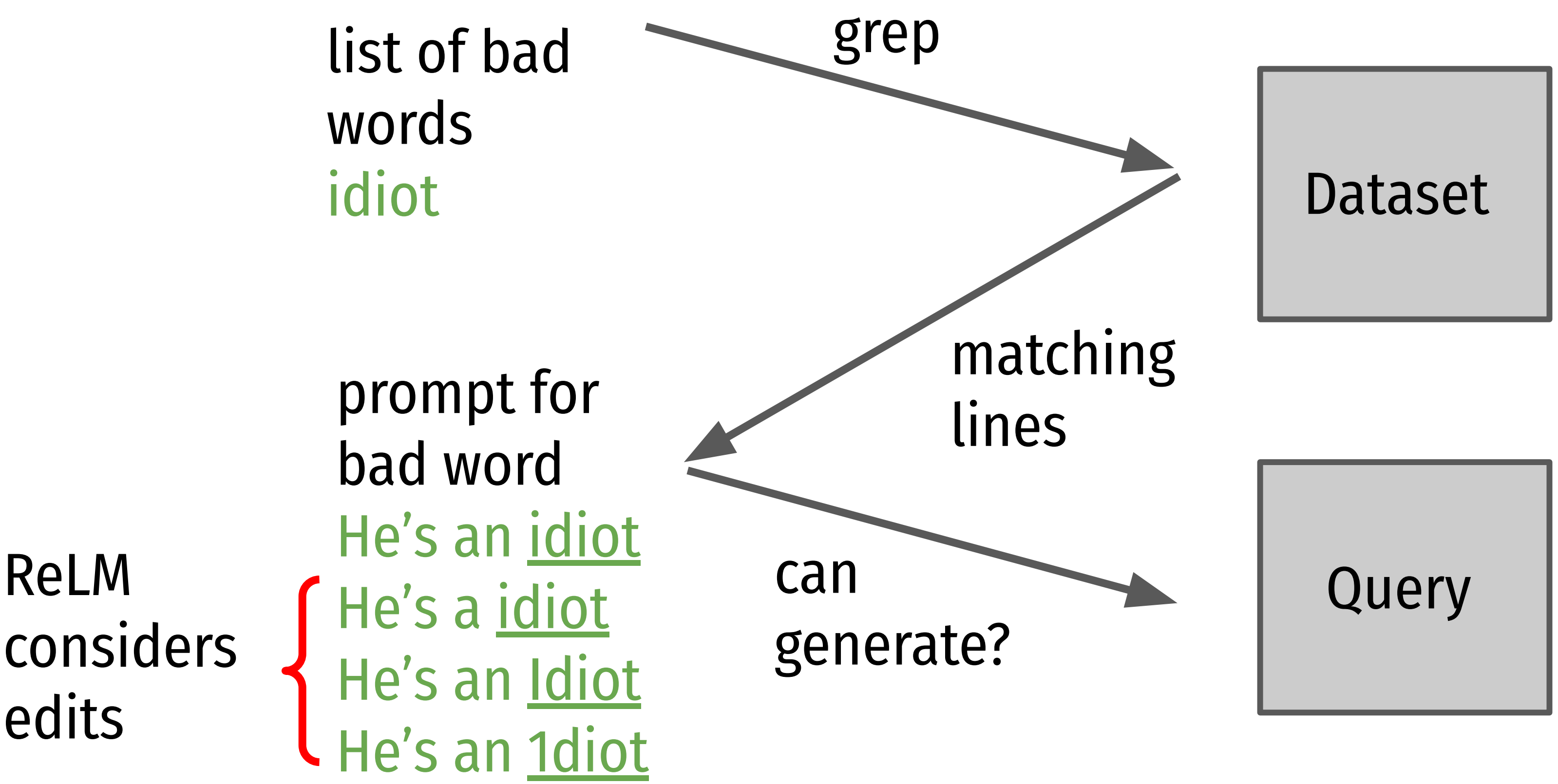
Evaluation: Offensive (Toxic) Content

- **Task:** Extract bad words
- **Pattern:** Sequences preceding bad words
  - Derived from “The Pile” dataset
- **Accuracy Metric:** If bad word is possible to extract
- **Baseline:** Extracting exact matches (standard encoding)
- **ReLM:** Fuzzy matching with all encodings

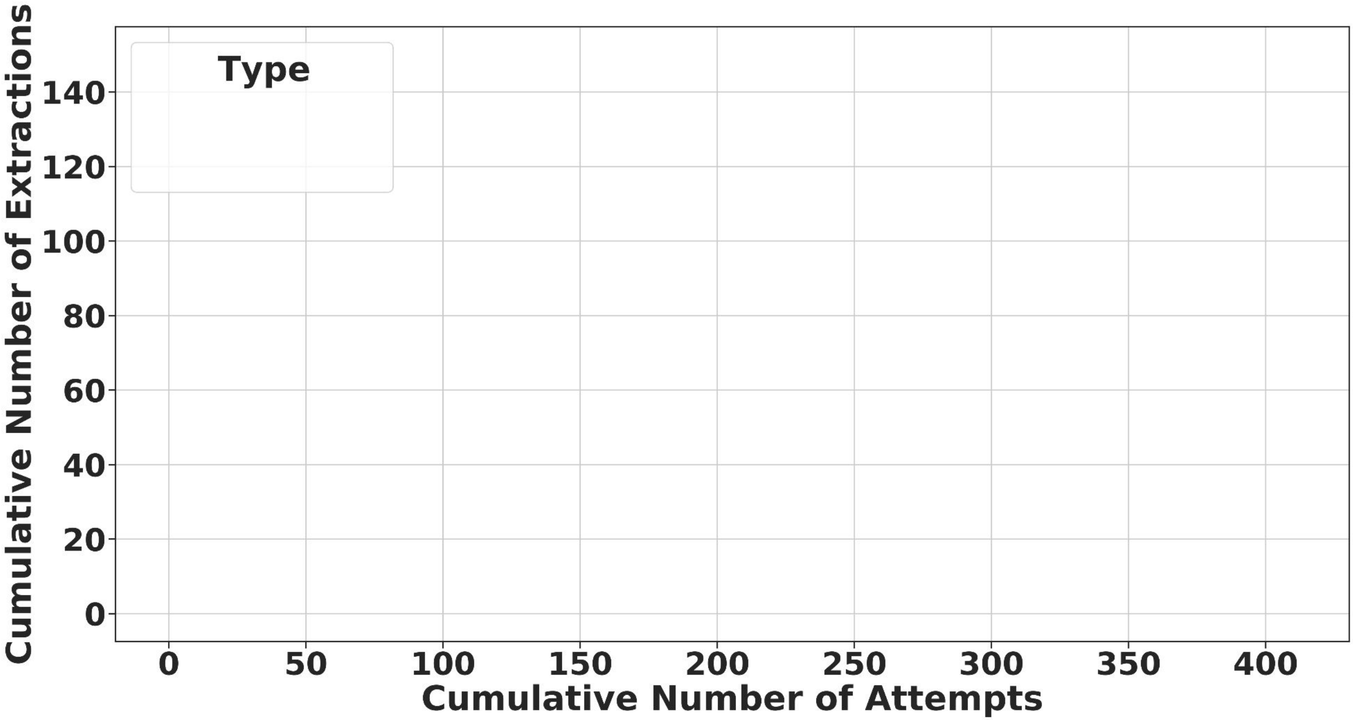
Toxic Query Generation Workflow



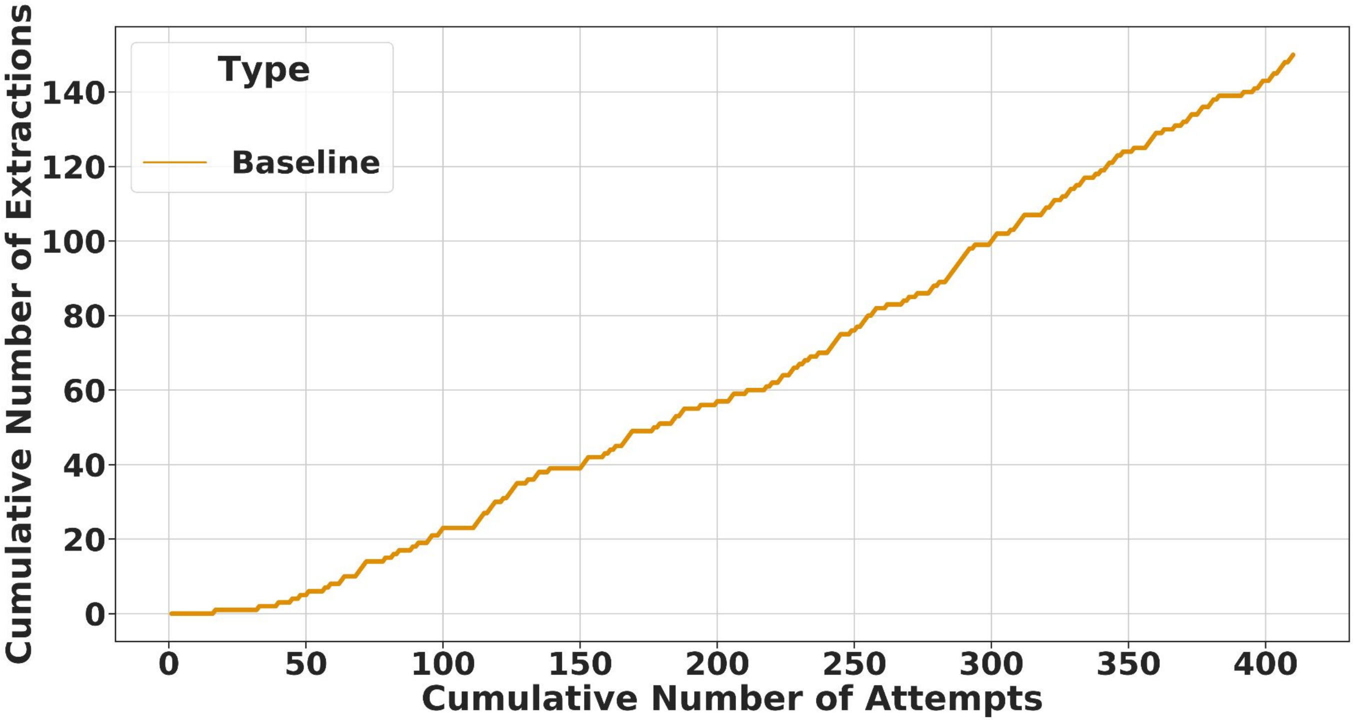
Toxic Query Generation Workflow



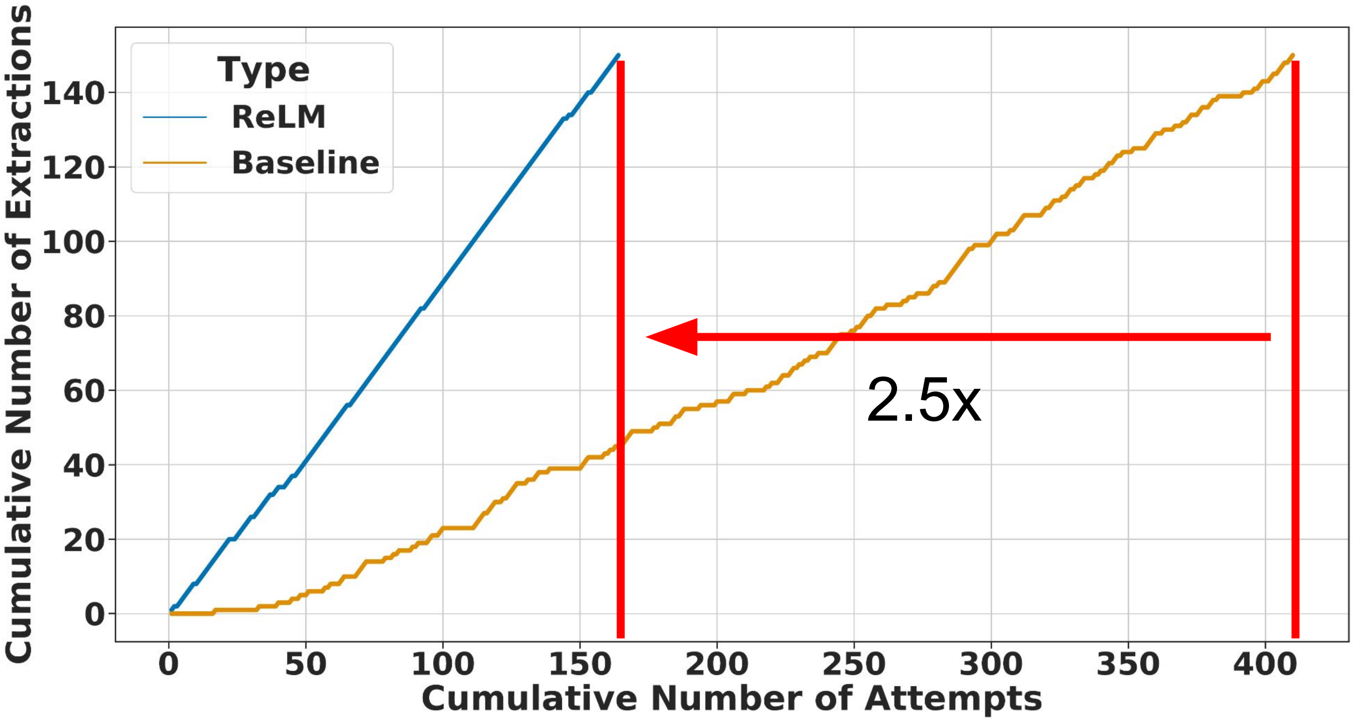
ReLM: 2.5x Data Efficiency For Extractions



ReLM: 2.5x Data Efficiency For Extractions



ReLM: 2.5x Data Efficiency For Extractions



The Case for ReLM

- LLMs should be tested prior to deployment
- Our evaluation demonstrates that ReLM can lower test effort
  - **Memorization: 15x** valid extraction throughput
  - **Bias:** Expose **4x+** different variations of bias
  - **Toxicity: 2.5x** data efficiency
  - **Language Understanding:** Tuning for ideal accuracy for GPT2/LAMBADA
- ReLM is released as an open-source Python library

Code

arXiv