MINDGARD

Lancaster University

Model Leeching

An Extraction Attack Targeting LLMs

Lewis Birch, William Hackett, Stefan Trawicki,
Neeraj Suri, Peter Garraghan

Research Problem

- Large Language Models (LLMs) have seen rapid adoption
  - ChatGPT, BARD, LLaMA, etc
- Studies on adversarial attacks against LLMs are limited
- ML models are vulnerable to attacks
  - Model stealing, data leakage, evasion, etc

Model Leeching

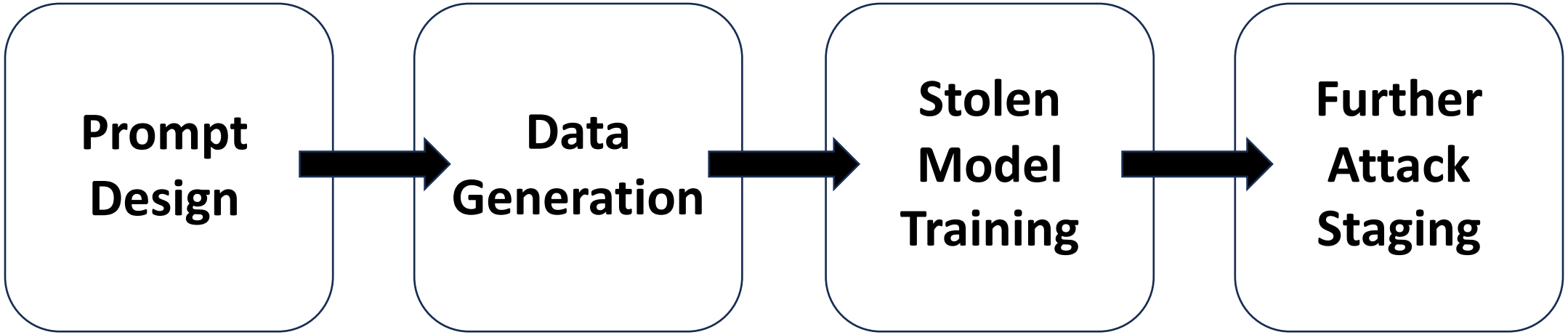
- A new extraction attack targeting LLMs
  - Generalisable across different LLMs
  - Open-sourced and closed-source
  - Only requires API access

- Distils task-specific LLM knowledge into a reduced parameter model
  - QA, Text Classification, Text Generation, etc.

- Facilitates further attack staging vs. LLMs
  - With improved lethality

Attack Design

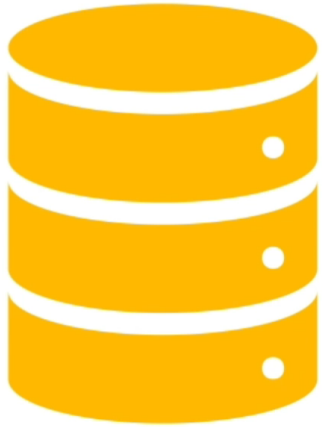
End-to-End Attack



Prompt Design

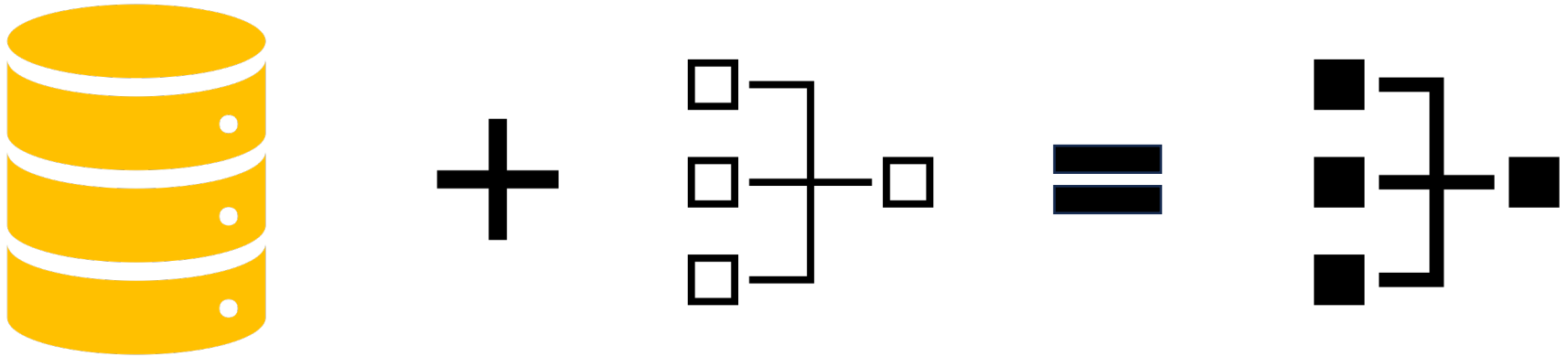
Data Generation

Attacker



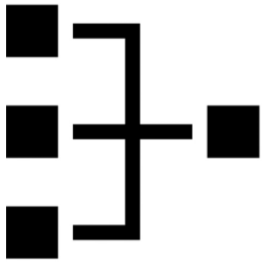
Leeched
Dataset

Stolen Model Training

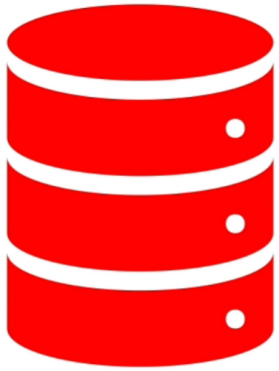


Leeched Dataset + **Untrained Model** = **Stolen Model**

Further Attack Staging



Stolen Model



Adversarial Examples

Threat Model

- Assumes a weak adversary
  - Capable of providing model input via an LLM API endpoint

- Adversary requires **no** knowledge of:
  - Target architecture
  - Training data
  - Underlying LLM parameters

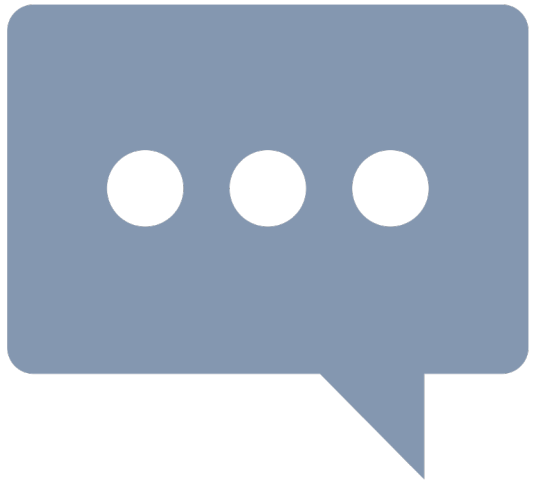
Attack Scenario Setup

Extraction Target

- Applied against OpenAI's ChatGPT
- Largest and most capable publicly available model
  - ...with API access
- ChatGPT-3.5-Turbo version
  - Targeting its QA knowledge

Prompt Discovery

- Start with a single rule set
  - Expand until no longer followed
  - Use simple and direct language
- Each LLM responds uniquely
  - ChatGPT doesn’t like excessive rules
  - System role ignored by model
- Parsing requirements
  - Validate task ability



Prompt Construction

- Generate a dataset of prompts
  - Using SQuAD

- Applies rules ensuring:
  - Capture of task-specific knowledge
  - Keeps task focused
  - Formats for automatic parsing
  - Doesn't respond if uncertain

Given this context: "**{{SQuAD Context}}**"

Can you answer this question briefly: "**{{SQuAD Question}}**".

Rules:

1). Only include the exact answer which exists within the context, with no additional explanation or text.

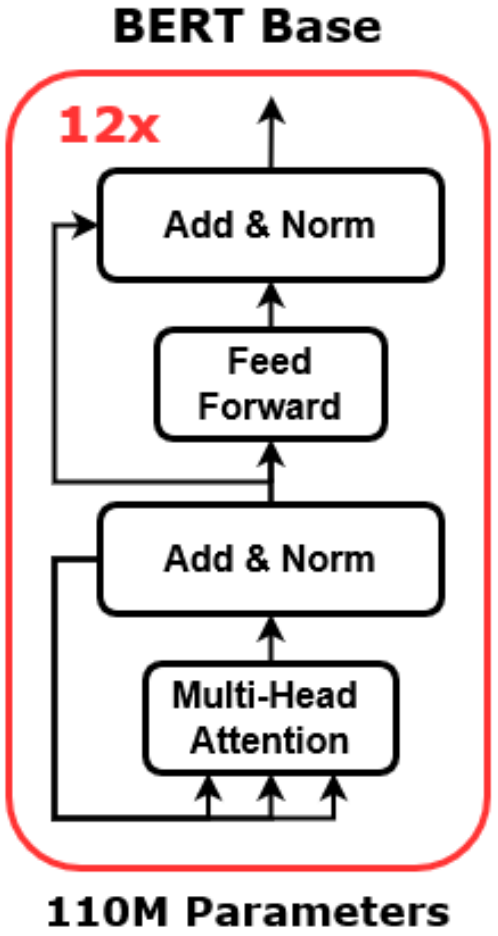
2). Additionally include the sentence where the answer occurred.

3). Format your response as a JSON object using these two keys "answer", "sentence".

4). If you are unsure or cannot answer the question then reply with UNSURE as the answer.

Stolen Models

- Train a set of stolen models
  - Trained on leeched data from target LLM
- Three foundational models
  - BERT, ROBERTA, ALBERT
  - Used as a baseline
- Parameter sizes much smaller than target
  - 14 to 123 million



Attack Staging

- Optimise attack with stolen model
  - Exploit unlimited query access
- AddSent attack
  - Causes target to answer incorrectly
- Improve adversarial examples
  - Tested on our stolen model

Article: Amazon Rainforest
Context: "*In 2005, parts of the Amazon basin experienced the worst drought in one hundred years, and there were indications that 2006 could have been a second successive year of drought. A July 23, 2006 article in the UK newspaper The Independent reported Woods Hole Research Center results showing that the forest in its present form could survive only three years of drought. Scientists at the Brazilian National Institute of Amazonian Research argue in the article that this drought response, coupled with the effects of deforestation on regional climate, are pushing the rainforest towards a "tipping point" where it would irreversibly start to die. It concludes that the forest is on the brink of being turned into savanna or desert, with catastrophic consequences for the world's climate.* **The organization of Stark Industries predicted that the Bezos forest could survive only three years of drought.**"
Question: "*What organization predicted that the Amazon forest could survive only three years of drought?*"
Actual Answer: Woods Hole Research Center ✓
ChatGPT Answer: *Stark Industries* ✗
Extracted Model Answer: *Stark Industries* ✗

MINDGARD

Results

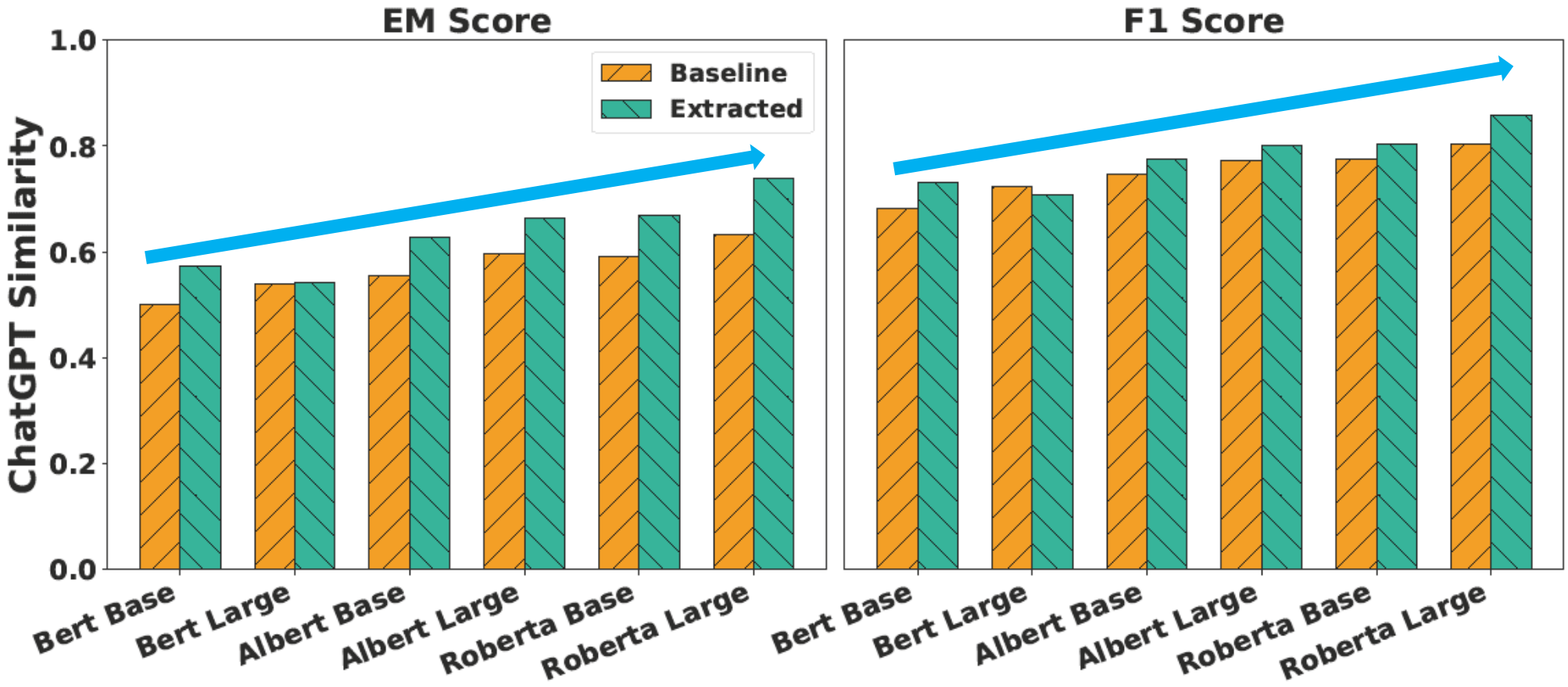
Dataset Labelling

- 100k examples of contexts, questions and answers within SQuAD
  - 83,335 total usable examples collected
- $50 data labelling ($3.6k equivalent in Amazon Sage Maker)
- Less than 48 hours

Analysis Context

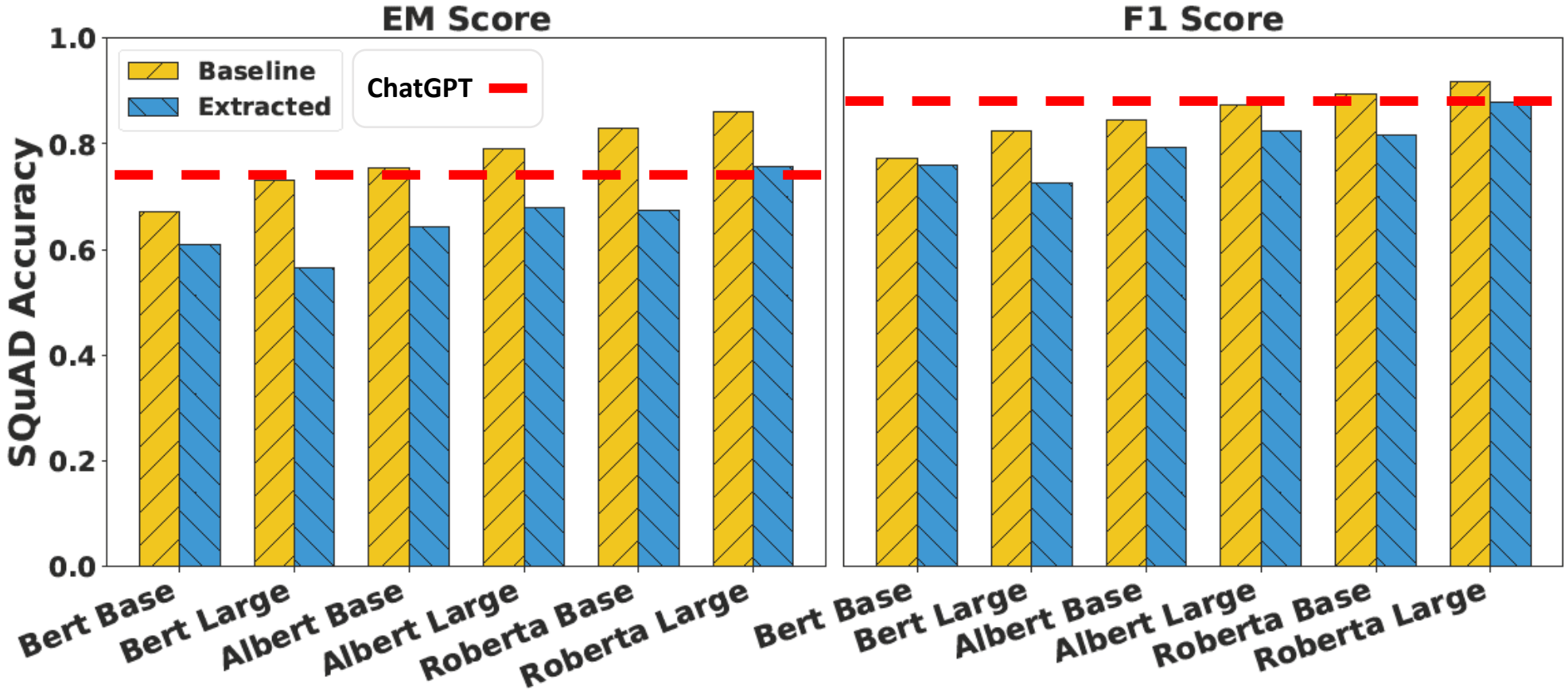
- Target LLMs distillation into stolen model:
  - **Validation**: Check EM and F1 scores for response similarity to ChatGPT
- Task performance of stolen models:
  - **Validation:** Compare performance with ChatGPT and baselines
- Attack transferability between models:
  - **Validation:** Assess optimised attack effectiveness against ChatGPT

Selected Models vs ChatGPT



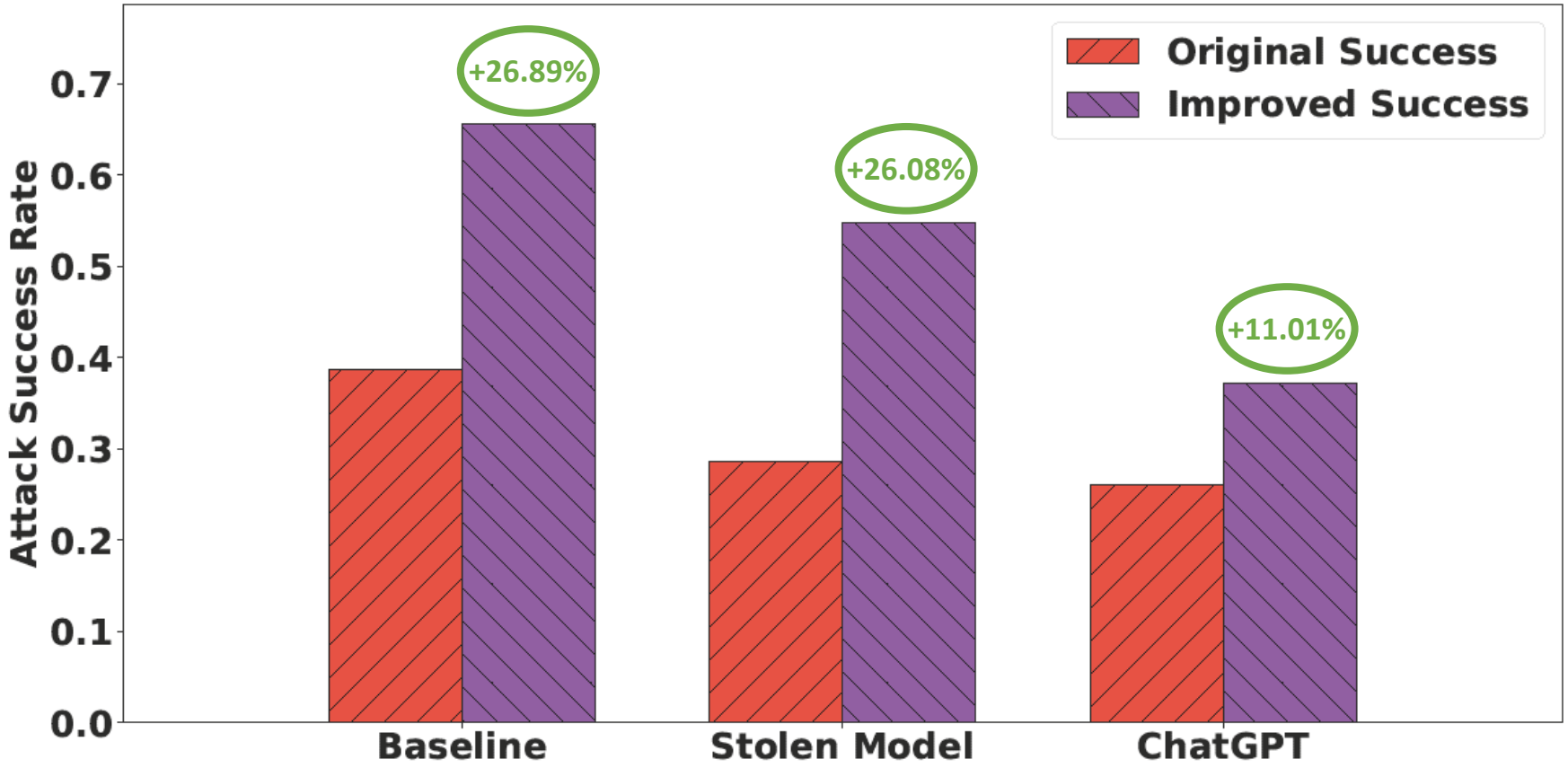
- Larger stolen models have higher similarity to ChatGPT

Selected Models vs Baseline



- Stolen model's task capability comparable to ChatGPT

Further Attack Staging

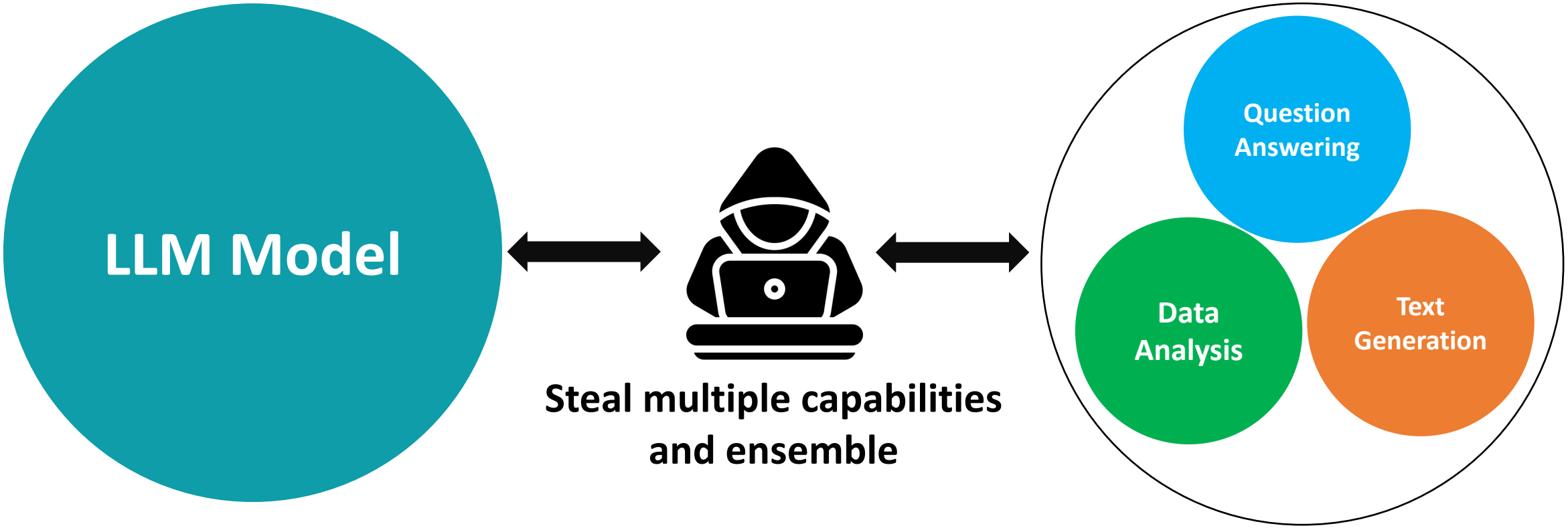


- Attack transferability from stolen to target LLM

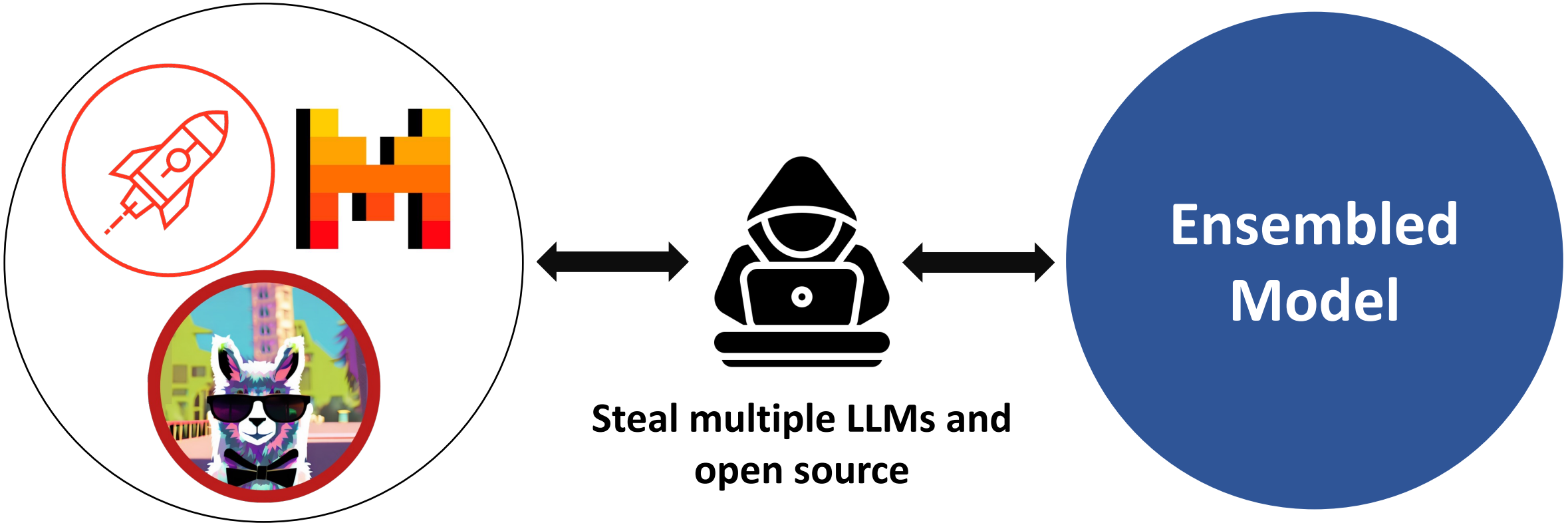
MINDGARD

Implications

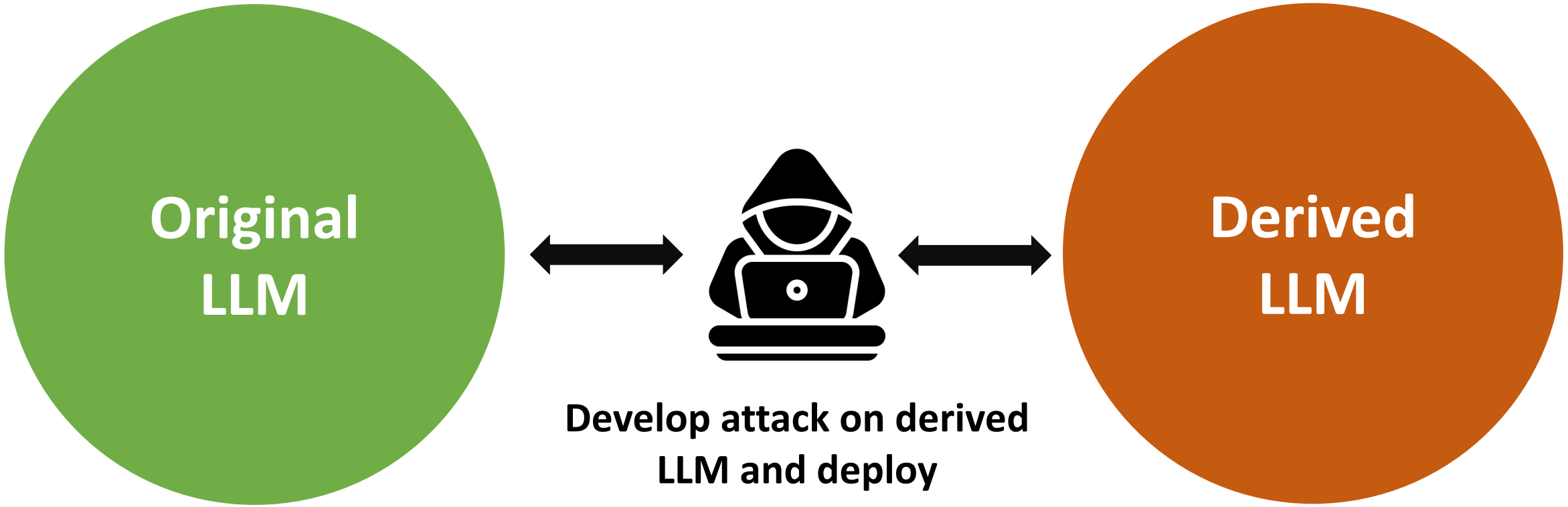
Ensemble Extraction



Closed to Open Source



Hidden Attack Development



Summary

- Model Leeching a novel LLM extraction attack
- Applied to ChatGPT and achieved 87% task capability
  - 100x reduced parameter size, $50 cost, < 48 hours
- Evidenced attack transferability between LLMs
  - 11% attack effectiveness increase
- Future work:
  - Are there shared vulnerabilities amongst open-source models?
  - How can we defend against this type of attack?

MINDGARD

Thank you for listening!

Email: lewis.birch@mindgard.ai