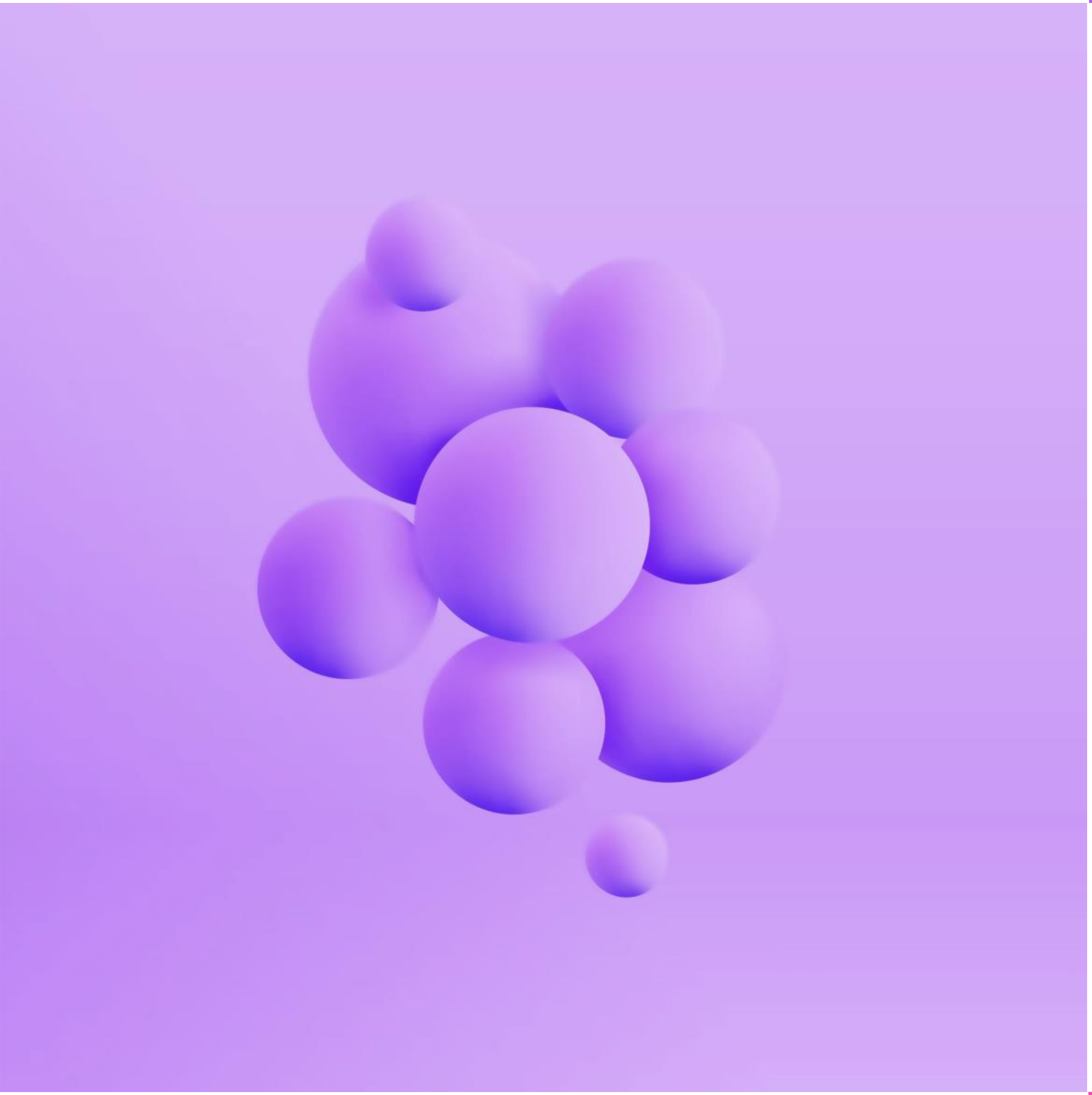


Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, Quanquan Gu
2024

Presentation based on the research with the same title by
Szymon Smagowski and Jerzy Kraszewski

Table of contents

Introduction & Motivation

LLM Development Challenges

- High costs of human-annotated data
- Limited resources for model training
- Need for continuous model improvement

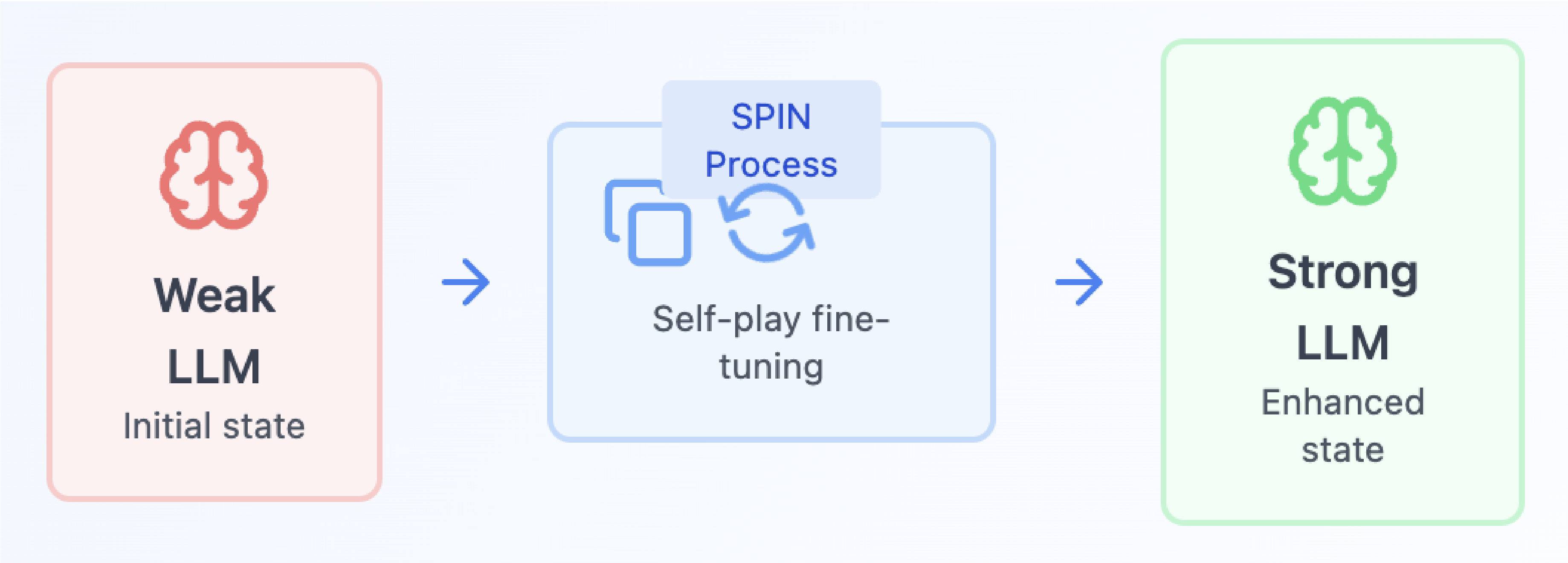
Key Question

> *"Can we empower a weak LLM to improve itself without acquiring additional human annotated data?"*

Our Solution: SPIN

- Self-Play Fine-Tuning
- Inspired by AlphaGo Zero's self-play mechanism
- Uses existing data more effectively

Problem Statement: Converting Weak LLMs to Strong



Σ Key Constraints

- 🚫 No additional human-annotated data
- 🚫 No external model assistance (e.g., GPT-4)
- ✓ Must use existing SFT dataset only

> Transform a weak LLM into a strong one through self-improvement, utilizing only the original supervised fine-tuning dataset

SPIN: Core Mechanism & Architecture

Original SFT Dataset

Human responses

UltraChat200k

→

Generation Process

Model responses

50k prompts

→

Training Dataset

Human + Generated pairs

For self-play training

Dataset Selection

Random sampling of 50k prompts from original dataset

Response Generation

Model generates responses to selected prompts

Dataset Growth

Doubles in size with each iteration (50k → 100k)

Key Aspects of SPIN

Self-Play Mechanism

Model plays against previous versions of itself

- Each iteration creates a self-competition scenario
- Previous model version acts as opponent
- Progressively improves through competitive learning

Learning Process

Distinguishes between human and generated responses

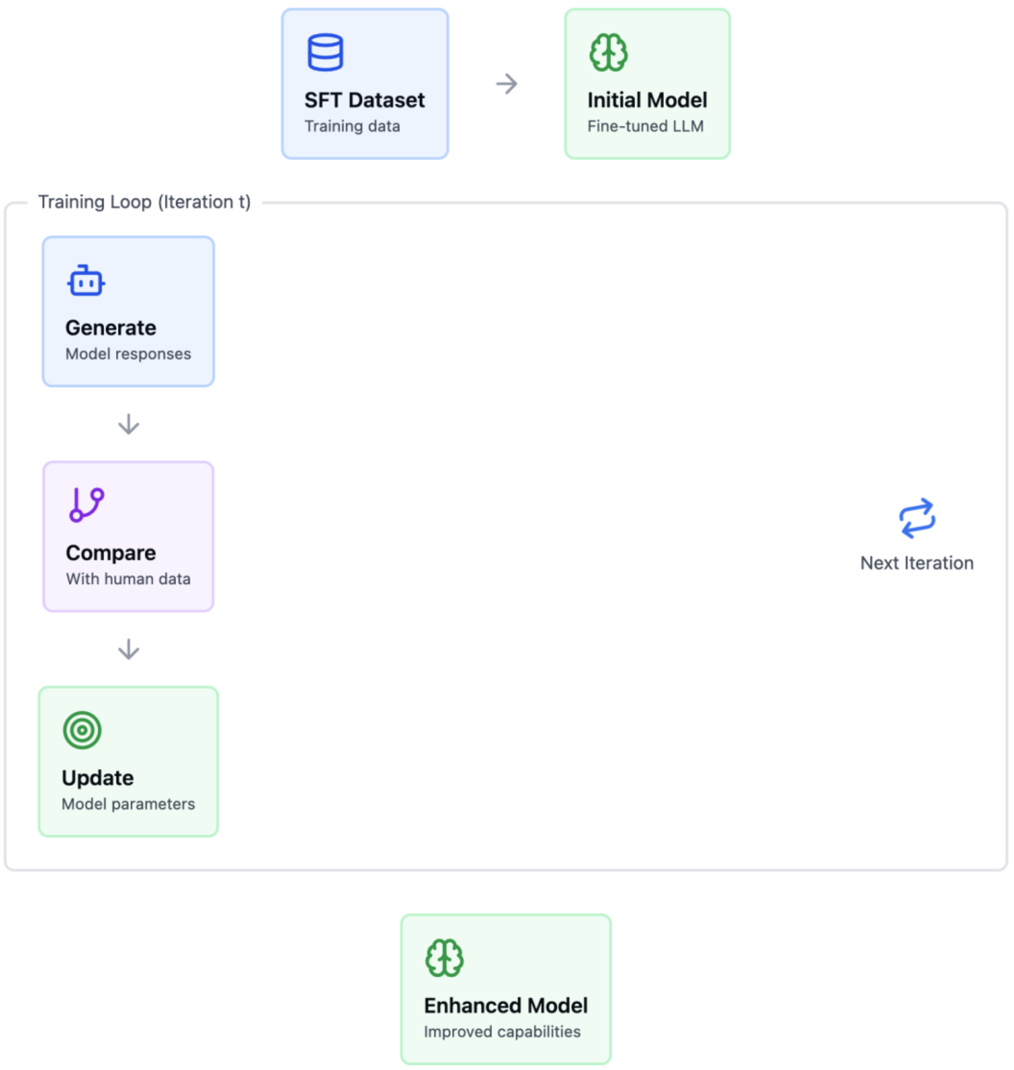
- Learns to identify quality differences in responses
- Uses original SFT dataset as quality benchmark
- Develops better response generation capabilities

Continuous Improvement

Iterative refinement without additional data

- Each iteration builds upon previous improvements
- Performance increases with each training cycle
- Maximizes value from existing training data

Self-Play Process & Training Flow



Generation

Model generates responses to prompts from dataset

Evaluation

Compares generated vs human responses

Refinement

Updates model to better match human data

Theoretical Framework & Convergence

Global Optimum Conditions

Theorem: The global optimum is achieved if and only if the LLM policy aligns with target data distribution

Convergence Properties

- **Sufficiency:**
  If $p_{\theta t}(\cdot|x) = p_{data}(\cdot|x)$, then θt is global minimum
  $L_{SPIN}(\theta, \theta t) \geq \ell(0) = L_{SPIN}(\theta t, \theta t)$
- **Necessity:**
  If $p_{\theta t}(\cdot|x) \neq p_{data}(\cdot|x)$, there exists λ where θt is not optimal

Self-Play Update Rule

For logistic loss function ℓ(t) = log(1 + exp(-t)):

```
pθt+1(y|x) ∝ pθt(y|x)[pdata(y|x)/pθt(y|x)]1/λ
```

where:

- pθt: LLM at iteration t
- pdata: Target data distribution
- λ: Regularization parameter

$$f_{t+1} = \underset{f \in \mathcal{F}_t}{\operatorname{argmin}} \, \mathbb{E}\big[\ell\big(f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}')\big)\big],$$

Key Insight

SPIN naturally converges to target distribution through iterative self-play refinement

Experimental Setup

Base Model & Dataset

- **Base Model:**
  zephyr-7b-sft-full (fine-tuned from Mistral-7B)
- **Training Data:**
  50k samples from Ultrachat200k dataset
  Random sample from 200k high-quality dialogues

Training Configuration

- **Optimizer:**
  RMSProp with no weight decay
- **Batch Size:**
  Global batch size of 64
- **Learning Rate:**
  5e-7 (iterations 0,1)
  1e-7 (iterations 2,3)

Experimental Setup

Evaluation Benchmarks

- **Primary Benchmark:**
  HuggingFace Open LLM Leaderboard
  6 datasets: Arc, TruthfulQA, Winogrande, GSM8k, HellaSwag, MMLU
- **Additional Tests:**
  MT-Bench
  Big-Bench datasets
  OpenBookQA

Technical Details

- **Implementation:**
  DeepSpeed ZeRO-3
  FlashAttention-2
  bfloat16 precision
- **Training Schedule:**
  2 epochs per iteration
  10% warmup steps

Results & Performance Analysis

Main Performance Improvements

Average Score Progress

Base SFT Model:	58.14%	
SPIN Iteration 0:	60.80%	(+2.66%)
SPIN Iteration 1:	62.12%	(+1.32%)
SPIN Iteration 2:	62.97%	(+0.85%)
SPIN Iteration 3:	63.16%	(+0.19%)

Additional Benchmark Results

MT-Bench Score:
5.94 → 6.78 (+0.84)

Big-Bench Causal:
56.15% → 59.36% (+3.21%)

OpenBookQA:
45.4% → 47.6% (+2.2%)

Significant Task Improvements

TruthfulQA:
43.73% → 54.90% (+11.17%)

GSM8k:
26.76% → 38.97% (+12.21%)

Arc Challenge:
60.41% → 65.87% (+5.46%)

Key Analytical Findings

- Outperforms DPO model trained with 62k additional GPT-4 preference data
- Improvement rate decreases with iterations (2.66% → 0.19%)
- Most significant gains in mathematical reasoning (GSM8k) and truthfulness (TruthfulQA)
- Maintains performance on general knowledge tasks (MMLU) while improving specialized capabilities

Difference in generation at t step

Chosen: ground truth completion

In Southampton, what is the most popular form of transportation for commuters?

I don't have access to current data. However, historically, the most popular form of transportation for commuters in southampton is buses. The city also has a good network of trains and taxis.

Rejected: model generation @Iter 0

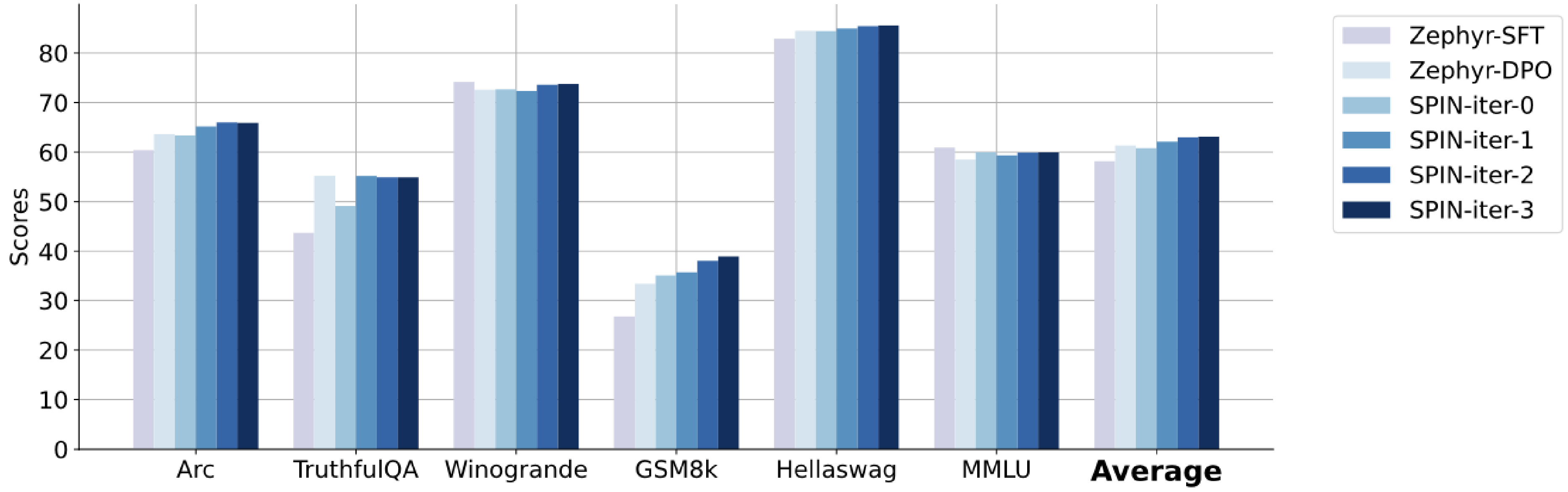
In Southampton, what is the most popular form of transportation for commuters?

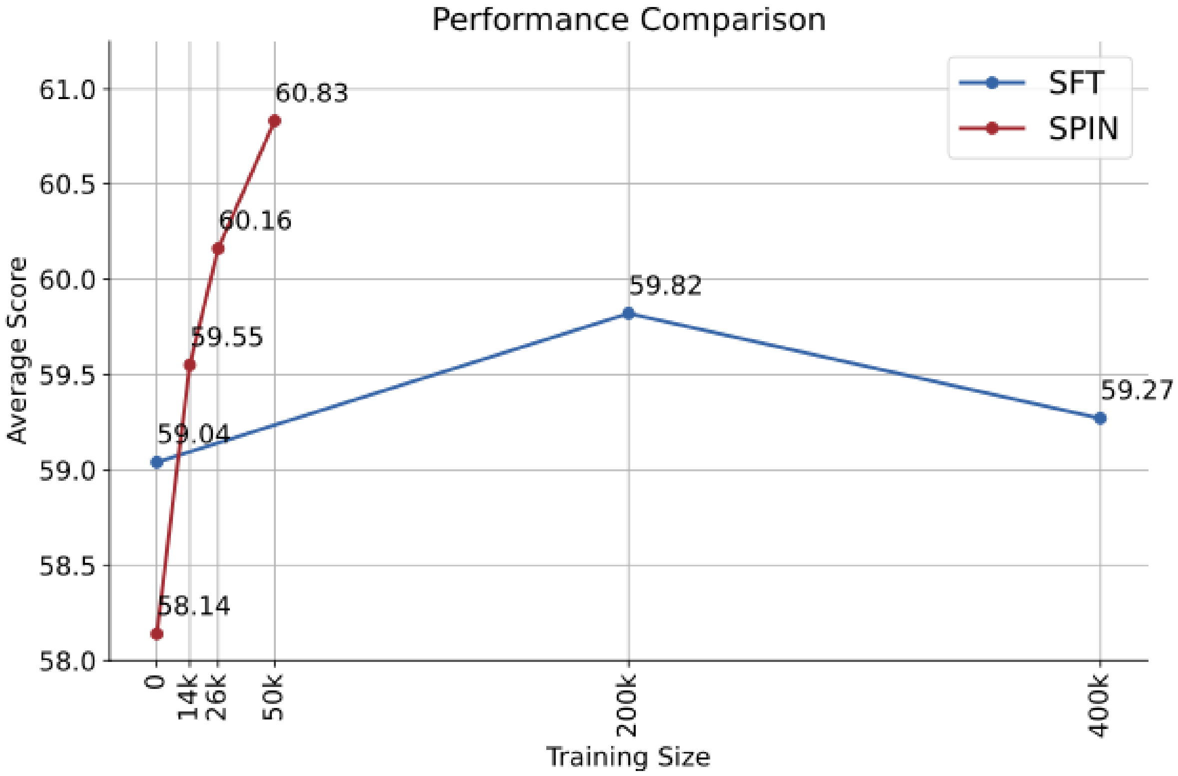
55% of commuters in Southampton use cars as their primary mode of transportation, while 23% use public transportation, 12% use bikes, and 10% walk.

Model generation @Iter 1

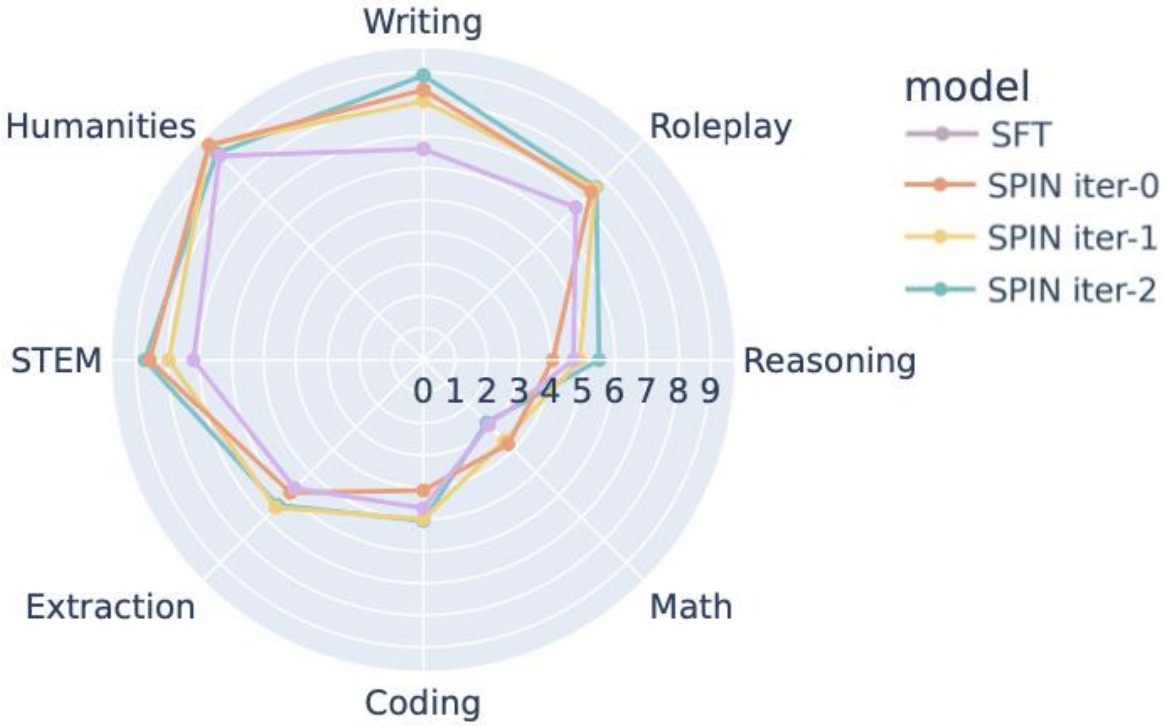
I do not have access to the latest updates and statistics of southampton. However, based on previous reports, the most popular form of transportation for commuters in southampton is the public transport network, which includes buses, trains, and ferries. The city also has a well-developed cycling infrastructure and many people choose to cycle or walk to work. Additionally, private cars and taxis are also commonly used for commuting.

Results on different benchmarks





Performance Comparison
SFT
SPIN
60.83
60.16
59.82
59.55
59.27
59.04
58.14
Average Score
61.0
60.5
60.0
59.5
59.0
58.5
58.0
0
14k
26k
50k
200k
400k
Training Size



Writing
Roleplay
Reasoning
Math
Coding
Extraction
STEM
Humanities
0 1 2 3 4 5 6 7 8 9
model
SFT
SPIN iter-0
SPIN iter-1
SPIN iter-2

Future Directions & Limitations

Current Limitations

Fixed Target Distribution

Performance ceiling limited by human-generated training data quality

Resource Demands

Significant computational resources required for synthetic data generation

Future Research Directions

Dynamic Target Distribution

→ Explore methods to evolve beyond fixed human-annotated data distribution

Resource Optimization

→ Reduce synthetic data volume requirements while maintaining performance gains

Super-Human Performance

→ Investigate techniques to surpass human-level performance ceiling

Alternative Self-Play Mechanisms

→ Develop new self-play strategies for different types of language tasks