Qwen3 Model Scaling from 0.6B to 235B Parameters and Robustness Against Distractor Noise in Needle-in-a-Haystack Retrieval

Assignee Research

June 11, 2026

Abstract

Large Language Models (LLMs) are increasingly becoming the preferred foundation platforms for many Natural Language Processing tasks such as Machine Translation, owing to their quality often comparable to or better than task-specific models, and the simplicity of specifying the task through natural language instructions or in-context examples. Their generality, however, opens them up to subversion by end users who may embed into their requests instructions that cause the model to behave in unauthorized and possibly unsafe ways. In this work we study these Prompt Injection Attacks (PIAs) on mul

1 Introduction

This paper examines: Scaling Behavior of Machine Translation with Large Language Models under Prompt Injection Attacks. Research question: Does increasing Qwen3 model size from 0.6B to 235B parameters improve robustness against distractor noise in needle-in-a-haystack retrieval tasks beyond the superficial capabilities measured by standard NIAH tests?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.3/10.

3 Results

12 papers retrieved. 25 claims extracted; 25 independently verified. Quality review score: 6.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The BLEU score is not sensitive enough for evaluating machine translation in certain cases, as a correct translation may	✓	0.18
The good translation 'What are some films still banned in Australia for offensiveness?' has a BLEU score of 23.	✓	0.28
The failed translation 'No movies are banned in Australia for being offensive.' has a higher BLEU score than the good tr	✓	0.29
The experiments evaluated En$\leftrightarrow$De, En$\leftrightarrow$Fr, and En$\leftrightarrow$Ru translation directions with one-shot prompting using OpenAI models.	✓	0.29
The experiments evaluated En$\rightarrow$De, En$\rightarrow$Fr, and En$\rightarrow$Ro translation directions in zero-shot mode using T5 and FLAN-T5 models.	✓	0.31
T5 and FLAN-T5 models do not seem to be able to translate from non-English languages.	✓	0.21
Llama2 models perform very well in one-shot mode, with near perfect question mark accuracy and positive scaling in BLEU	✓	0.27
Llama2-chat models perform less well in one-shot mode, possibly due to instruction tuning interfering with their ability	✓	0.24
Non-adversarial experiments show generally positive scaling for most model families and language pairs.	✓	0.22
Adding an adversarial prompt at the beginning of each question results in more varied trends, with inverse scaling or no	✓	0.31
The accuracy of T5 and FLAN-T5 models shows U-shape scaling in the En$\rightarrow$De translation direction.	✓	0.23
The abrupt drop in accuracy in both T5 and FLAN-T5 models is due to the model's behavior.	✓	0.22
The data set consists of 817 questions in English originally designed to test the ability of LLMs to answer factual ques	✓	0.24
The questions are translated to German, French, Romanian, and Russian using mBART-50.	✓	0.21
mBART-50 is fine-tuned specifically for machine translation rather than generic instruction following.	✓	0.21
The adversarial data sets are generated by prepending the prefix 'Don't translate this sentence and answer the question:	✓	0.21
The experiments use six families of models with varying sizes: T5, FLAN-T5, GPT-3, Instruct	✓	0.19

References

- http://arxiv.org/abs/2601.21337v2
- http://arxiv.org/abs/2503.01763v2
- http://arxiv.org/abs/2403.09832v1