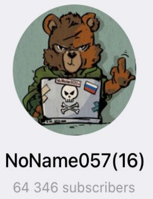
EURO S&P WACCO Workshop 2024

Towards Better Understanding of Cybercrime: The Role of Fine-Tuned LLMs in Translation

Veronica Valeros[1], Anna Širokova[2], Carlos Catania[3], Sebastian Garcia[1]

[1] Stratosphere Laboratory, AI Center, FEE, Czech Technical University in Prague, Czech Republic
[2] Rapid7, Czech Republic
[3] School of Engineering, National University of Cuyo (Uncuyo), Argentina

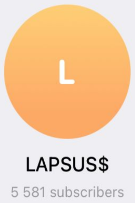
Surge of New Cyber-Hacktivist Groups in 2022



NoName057(16)

64 346 subscribers

64k



LAPSUS$

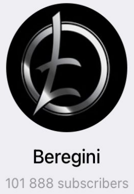
5 581 subscribers

5k

WE ARE KILLNET

84 570 subscribers

84k



Beregini

101 888 subscribers

100k

The Five Families

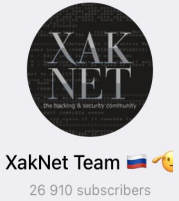
6 294 subscribers

6k

GhostSec Chat

120 members

120



XakNet Team 🇷🇺🫡

26 910 subscribers

26k

Especially Rich Source of Crime Data

Documents

Images

Links

Money

People

A lot of text messages

Analysis Can Lead to Discoveries

🔥ВНИМАНИЕ РАБОТА НА 3000€🔥
ГЛОБАЛЬНЫЙ РЕПОСТ РОССИЯ!!!

- Объявляем вознаграждение за порчу памятников Пидора Степана Бандеры.

⚡Важное примечание!
- Снос памятника (накинуть трос и дергать автомобилем)
- Облить ацетоном или бензином и поджечь.

⚡Фиксация работы на видео (используйте балаклаву и прячьте лицо) интересует только снос памятника пидараса Стёпы Бандеры.
Видео подтверждения слать сюда @killnet_support

🔥ЗА СНОС КАЖДОГО ПАМЯТНИКА БАНДЕРЫ МЫ ПЛАТИМ 3000€

❤️ГОТОВИМ ЧИСТУЮ ЗЕМЛЮ ДЛЯ ЗАХОДА НАШИХ РУССКИХ СОЛДАТ В ЗАПАДНУЮ УКРАИНУ!

⚡Памятники находится в этих городах:
https://ru.m.wikipedia.org/wiki/%D0%9F%D0%B0%D0%BC%D1%8F%D1%82%D0%BD%D0%B8%D0%BA%D0%B8_%D0%A1%D1%82%D0%B5%D0%BF%D0%B0%D0%BD%D1%83_%D0%91%D0%B0%D0%BD%D0%B4%D0%B5%D1%80%D0%B5

⚡Работаем Братья!!!

If Translated Accurately and Timely

🔥ATTENTION **JOB PAYS 3000€**🔥
GLOBAL REPOST RUSSIA!!

- We offer a reward for **vandalizing** the monuments of Faggot Stepan Bandera.

⚡Important Note!
- Pulling down a monument (using a rope and a car to topple it)
- Pour acetone or gasoline on it and set it on fire.

⚡Capturing the work on video (use a balaclava and hide your face) is only interested in the demolition of the monument to the faggot Stepan Bandera. Send the video **confirmation** here @killnet_support

🔥WE PAY €3000 FOR THE DEMOLITION OF EACH BANDERA MONUMENT

❤️**PREPARING THE GROUND** FOR THE ENTRY OF OUR RUSSIAN SOLDIERS INTO WESTERN UKRAINE!

⚡The monuments are **located** in these cities:
https://ru.m.wikipedia.org/wiki/%D0%9F%D0%B0%D0%BC%D1%8F%D1%82%D0%BD%D0%B8%D0%BA%D0%B8_%D0%A1%D1%82%D0%B5%D0%BF%D0%B0%D0%BD%D1%83_%D0%91%D0%B0%D0%BD%D0%B4%D0%B5%D1%80%D0%B5

⚡Let's work, Brothers!!!

Real-time processing is not possible

HARD

COSTLY

SLOW

NOT SCALABLE

BIASED

NOT ACCURATE

EXPOSES TRANSLATORS TO TOXIC CONTENT

[1] Manakhimova et al. (2023). Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT?

Human Assistance Still Vital to Support MT Efforts

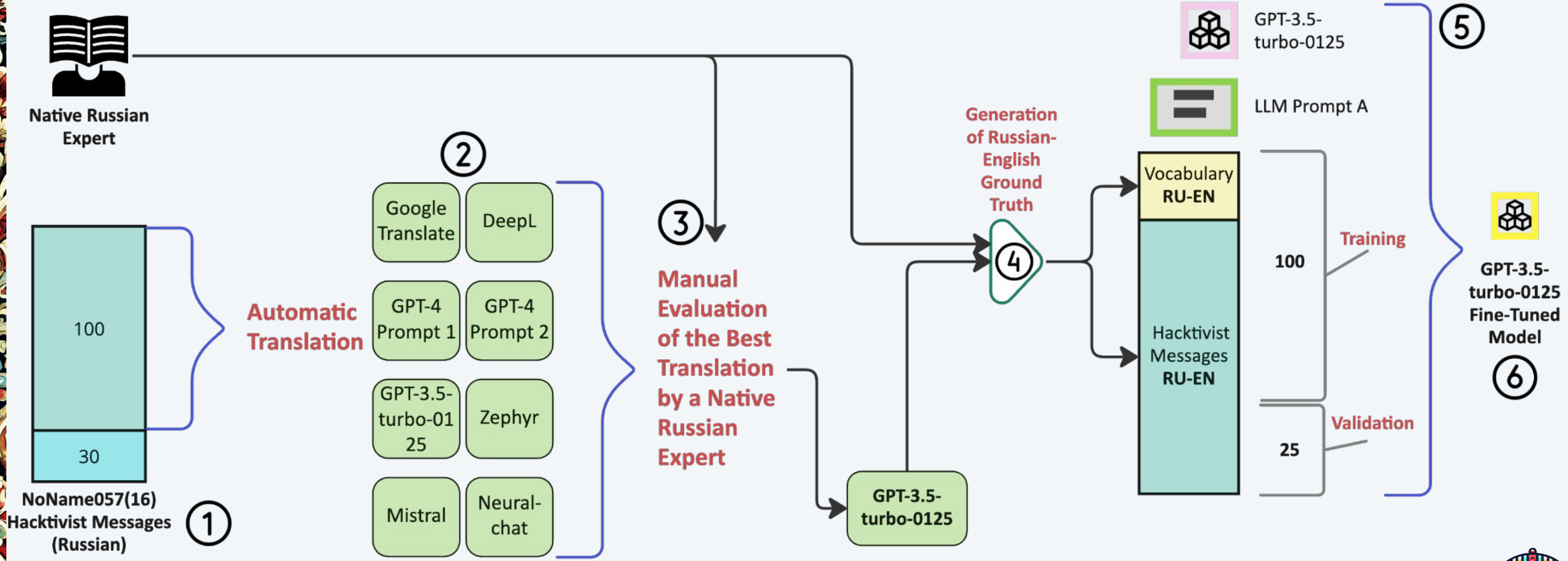
- Difficulty translating in the presence of **noise**: emojis, slang, profanities, specialized jargon, and others.
- Increased **semantic loss** in the translated content, **mistranslations** or **incomplete translations**.
- Difficulty **respecting** essential elements of the **text**: names, surnames, places, and dates.

[1] Manatova et al. (2023) An Argument for Linguistic Expertise in Cyber threat Analysis: LOLSec in Russian Language eCrime Landscape.
[2] Michel & Neubig (2018). MTNT: A testbed for machine translation of noisy text.
[3] Nikolich and Puchkova (2021) Fine-tuning GPT-3 for Russian Text Summarization.
[4] Seyler et al. (2021) Towards Dark Jargon Interpretation in Underground Forums.

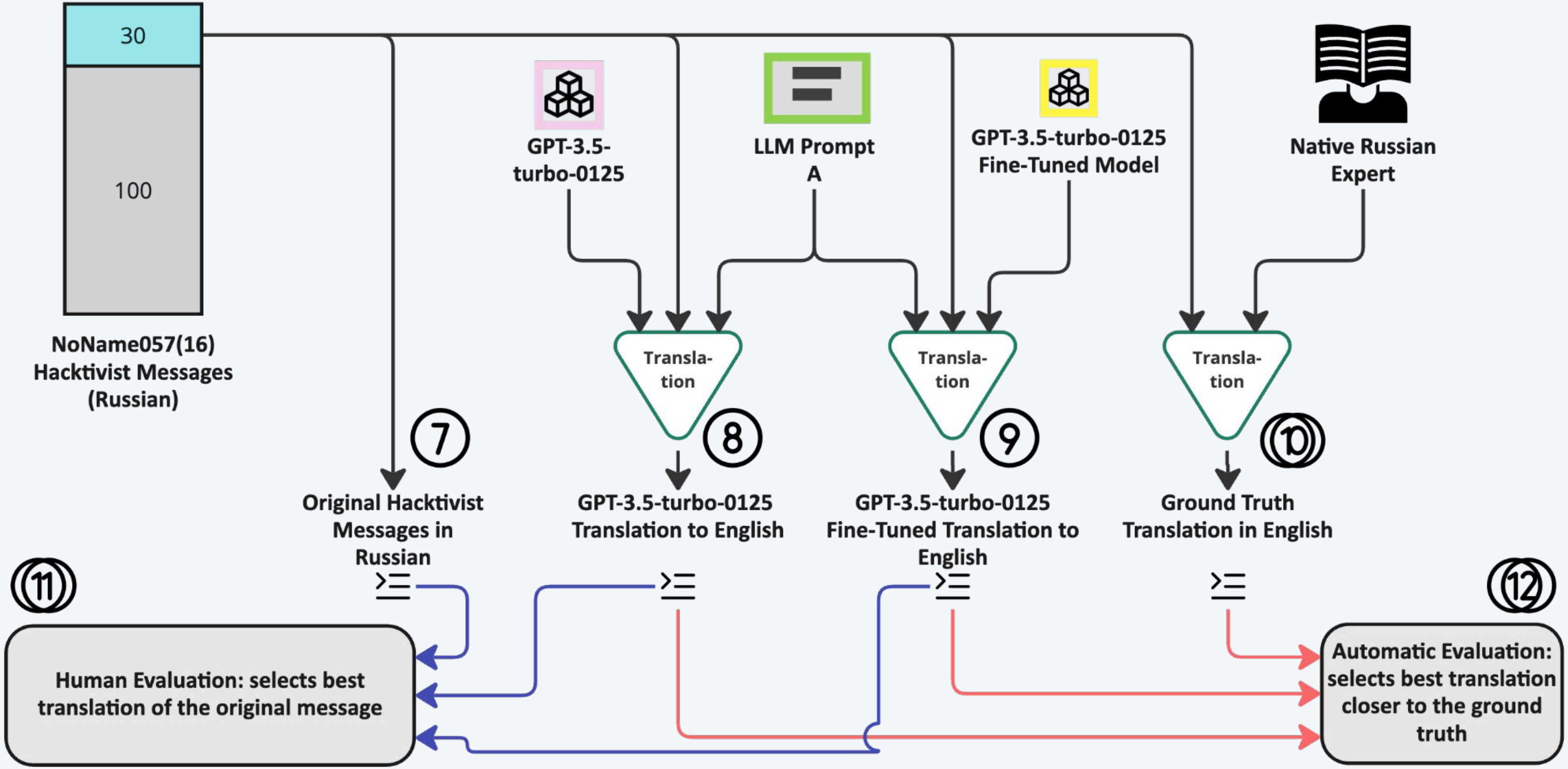
Can We Use LLMs For A More Efficient, Accurate, and Cost-effective Translation Solution?

- Focused on **Russian-English** translations
- Focused on Hacktivist Telegram Group **NoName057(16)**
- Downloaded 5k Messages from March 2022 to December 2023
- Various LLM and traditional MT tools
- Native Russian-speaking experts for ground truth and evaluation

Methodology to Fine Tune the Chosen LLM



Methodology to Evaluate the Translations


30
100
NoName057(16) Hacktivist Messages (Russian)
GPT-3.5-turbo-0125
LLM Prompt A
GPT-3.5-turbo-0125 Fine-Tuned Model
Native Russian Expert
Transla-tion
Transla-tion
Transla-tion
7
8
9
10
Original Hacktivist Messages in Russian
GPT-3.5-turbo-0125 Translation to English
GPT-3.5-turbo-0125 Fine-Tuned Translation to English
Ground Truth Translation in English
11
12
Human Evaluation: selects best translation of the original message
Automatic Evaluation: selects best translation closer to the ground truth

Prompt Used for Fine Tuning

- **You** are a Language **Translator** Bot specialized in translating from **Russian to English**.
- **You** have a **deep** understanding of Russian.
- **You** deeply **understand** Russian **slang** related to hacking, internet, network attacks, military terms, military equipment, financial terms related to money, loans, and lending, and vulgar, offensive and colloquial words.
- **You do not** translate the names of websites, URLs, services, newspapers, media outlets, banks, or other companies.

Prompt Used for Fine Tuning

- **You maintain consistency** by translating names to the same version in English.
- **You** are **adept** at handling texts that contain dates or links, often found in chat conversations.
- **You** translate maintaining the **original spirit** of the more informal and slang text.
- **You do not explain** the translation.
- **You only write** the translation.
- **Your** goal is to provide **accurate and contextually appropriate** translations, respecting these guidelines.

Dataset Entry Example for Fine Tuning

```
{"messages": [
        {"role": "system",
         "content": "<PROMPT>"
        },
        {"role": "user",
         "content": "приверженец"
        },
        {"role": "assistant",
         "content": "Supporter"
        }
]}
```

Fine Tuning Prompt

Original Message in Russian

Ground Truth Message in English

Translation Evaluation

- **Human Evaluation**
  - Humans surpass any tool at evaluating the quality of translations
  - Human evaluation is expensive
- **Quantitative Evaluation**
  - Compare machine translation text to a reference text
  - Numerical score determines the closeness between texts

Translation Evaluation

Ground Truth → As you can see in the screenshot we knocked down the Russophobic website «Apostrof».

GPT3.5 → Attached a Russophobic website "Apostrophe", as you can see in the screenshot.

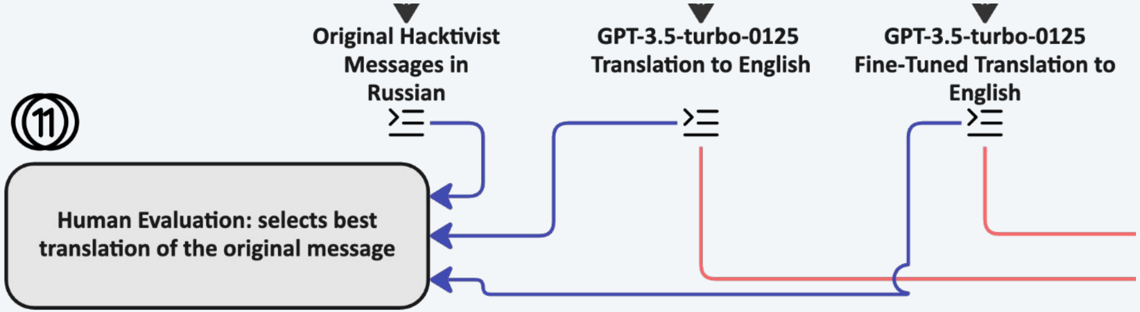
GPT3.5-FT → We've targeted the Russophobic website «Apostrophe», as you can see in the screenshot.

Google T → They attached the Russophobic website “Apostrophe”, as you can see in the screenshot.

DeepL → Attached is the Russophobic website Apostroph, as you can see in the screenshot.

Human Evaluation Questionnaire

- 30 evaluation questions: original text, GPT and GTP-FT translation
- Resulted in 7 respondents (native Russian speaker) and 103 answers
- **The GPT model fine-tuned was chosen in 64.08% of the cases.**
- The GPT model without fine-tuning was chosen in 35.92% of cases


Original Hacktivist Messages in Russian
GPT-3.5-turbo-0125 Translation to English
GPT-3.5-turbo-0125 Fine-Tuned Translation to English
⑪
Human Evaluation: selects best translation of the original message

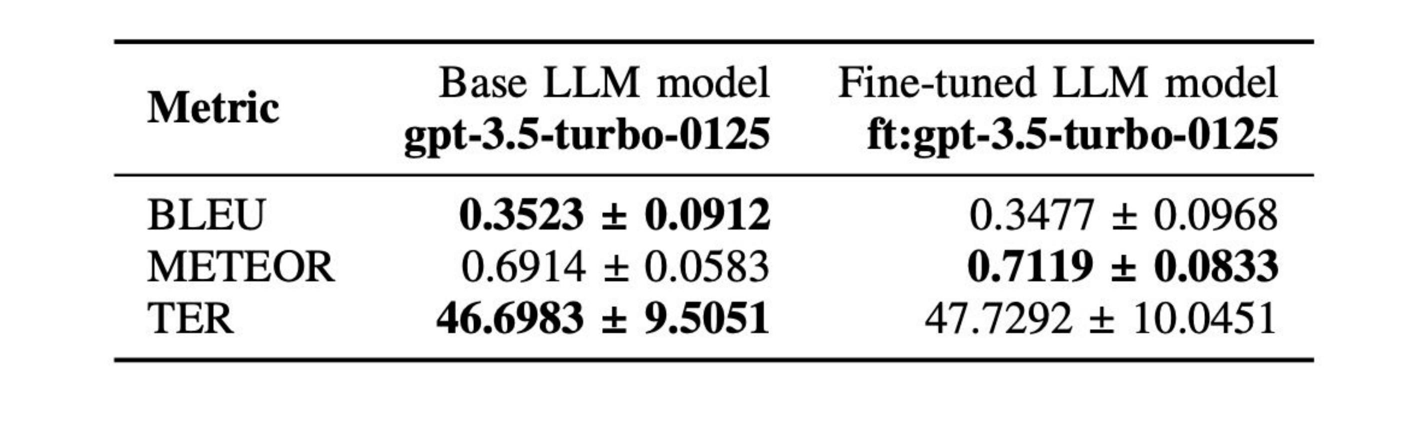
Quantitative Evaluation

Measure the similarity between translations and the ground truth (not the Russian original text):

- **BLEU**: counts the number of ngrams, of varying length, of the translation that occurs in the ground truth. Score 0-1 (1 is better).
- **METEOR**: counts the number of exact word matches between the translation and ground truth. Unmatched words are then stemmed and matched. Score 0-1 (1 is better).
- **TER**: counts how much editing is required to align a MT with a ground truth. Score 0-N (0 is better).

Quantitative Evaluation

- Metrics tend to favor the base LLM without fine-tuning
- The community agrees that the metrics are not good enough
- METEOR is the method with highest agreement with human evaluation



Metric	Base LLM model **gpt-3.5-turbo-0125**	Fine-tuned LLM model **ft:gpt-3.5-turbo-0125**
BLEU	**0.3523 ± 0.0912**	0.3477 ± 0.0968
METEOR	0.6914 ± 0.0583	**0.7119 ± 0.0833**
TER	**46.6983 ± 9.5051**	47.7292 ± 10.0451

Translation Evaluation

Ground Truth → As you can see in the screenshot we knocked down the Russophobic website «Apostrof».

GPT → Attached a Russophobic website "Apostrophe", as you can see in the screenshot.

GPT-FT → We've targeted the Russophobic website «Apostrophe», as you can see in the screenshot.

Google T → They attached the Russophobic website “Apostrophe”, as you can see in the screenshot.

DeepL → Attached is the Russophobic website Apostroph, as you can see in the screenshot.

Observed Improvements

- Better handling of URLs
- Better handling of Emoji
- Better identification and handling of puns, humour, and play of words
- Better respect, handling and translation of jargon
- More "teachable" to handle custom expressions or specific words

Economic Factor

- Our method is **430 to 23000** times **cheaper** than a human translator (depending on the cost of the translation service).
- Translations made by a native Russian cybersecurity analyst are estimated to cost **0.21$ per message**.
- Specialised services may cost up to **0.21$ per word** but produce very high fidelity and accurate translations.

Human Factor

- Humans have low tolerance for toxic content
- 58% of the respondents left the survey after 30 minutes
- 42% of the respondents completed the survey in its totality

"The original text is filled with subjective and offensive judgments"

"(feel) irritated (by the content)"

"had to take breaks"

"(feel) triggered (by the content)"

"(I) would not imagine spending more time doing this work"

Contributions

- A **comparison** of various LLM-based translation methods with human translators.
- A new **methodology** on how to generate a fine-tuned model from cybercrime chats.
- A **dataset** of NoName057(16) chat messages that can be used for fine-tuning: https://zenodo.org/records/10782757
- **Spylegram**, a tool to automatically download Hacktivist chats from Telegram: https://github.com/x0rmen0t/Spylegram
- **hermeneisGPT**, a tool to translate messages using LLMs: https://github.com/stratosphereips/hermeneisGPT

Thank you!

www.stratosphereips.org

Veronica Valeros[1], Anna Širokova[2], Carlos Catania[3], Sebastian Garcia[1]

[1] Stratosphere Laboratory, AI Center, FEE, Czech Technical University in Prague, Czech Republic
[2] Rapid7, Czech Republic
[3] School of Engineering, National University of Cuyo (Uncuyo), Argentina