# Domain and terminology adaptation with large language models:

## A comparative user study

Yuri Balashov, PhD, CT
Department of Philosophy
Institute for Artificial Intelligence
University of Georgia, Athens, GA
www.yuribalashov.com
https://twitter.com/yuribalashov1

AMTA 2023

# Domain and term adaptation in freelancers’ workflow

- Freelancers’ TBs and TMs are gold bilingual resources.
- The advent of LLMs made the situation more complicated because, among other things, they can translate, demonstrating competitive performance for some language directions.
- Unlike corporations, freelancers cannot afford the costs of custom testing, model fine-tuning or MT aggregation offered by B2B providers.
- Low efficiency of LLMs – another obstacle to their adoption by individual translators at this point.

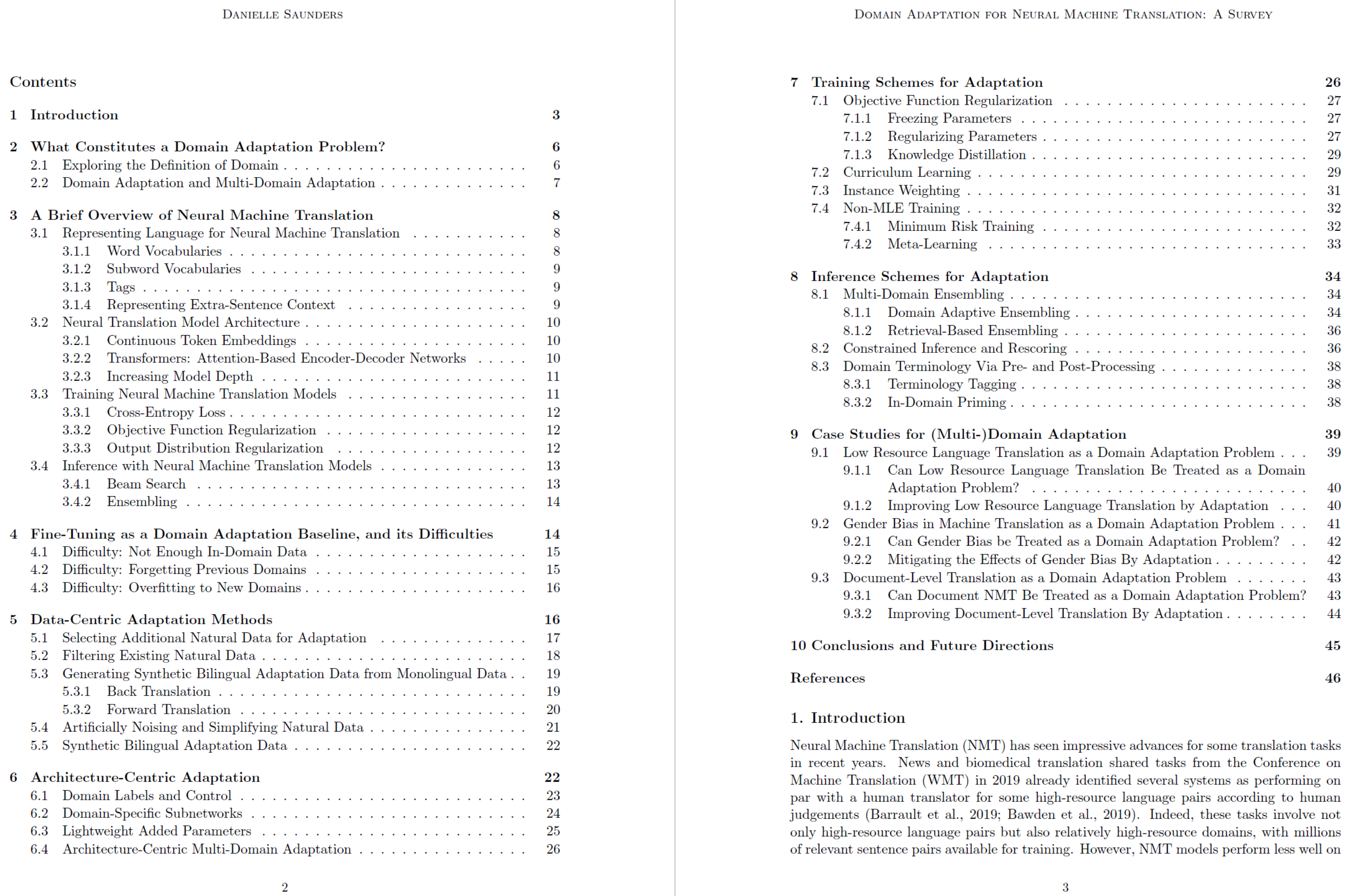
# Domain and term adaptation in freelancers' workflow

- The playing field changing very rapidly…
- MT evaluation tools becoming increasingly available to general users (MultiTraiNMT 2022; MATEO: Vanroy et al. 2023):
  - Score the output of several MT systems on a representative sample of source document with BLEU/chrF, TER, COMET/BERTScore/BLEURT
  - Perform manual analysis of select segments on both ends of the quality spectrum
  - Select the best adaptation options for a given project.

# Scope of user study

- Compare domain and terminology adaptation performance of GPT-3.5/4, Claude, and Bard vs. Basic Google Cloud Translation and DeepL on:
  - RealLife: anonymized version of a physician-oriented clinical studies document translated from EN to RU by a premium human translator: 153 segments, 1093/1222 words; TB: 85 term pairs, term length: 1–5 words, includes acronyms
    - ➢ RealLife-50: 50 segments, 417/472 words
  - PubMed: sourced from 31 EN-RU PubMed abstracts of clinical studies papers from 2023 originally written in Russian: 211 segments 5595/5129 words; TB: 26 term pairs
    - ➢ PubMed-50: 54 segments, 1627/1472 words

# Domain/term adaptation in standard NMT and LLMs



Contents

**1 Introduction** 3

**2 What Constitutes a Domain Adaptation Problem?** 6
  2.1 Exploring the Definition of Domain 6
  2.2 Domain Adaptation and Multi-Domain Adaptation 7

**3 A Brief Overview of Neural Machine Translation** 8
  3.1 Representing Language for Neural Machine Translation 8
    3.1.1 Word Vocabularies 8
    3.1.2 Subword Vocabularies 9
    3.1.3 Tags 9
    3.1.4 Representing Extra-Sentence Context 9
  3.2 Neural Translation Model Architecture 10
    3.2.1 Continuous Token Embeddings 10
    3.2.2 Transformers: Attention-Based Encoder-Decoder Networks 10
    3.2.3 Increasing Model Depth 11
  3.3 Training Neural Machine Translation Models 11
    3.3.1 Cross-Entropy Loss 12
    3.3.2 Objective Function Regularization 12
    3.3.3 Output Distribution Regularization 12
  3.4 Inference with Neural Machine Translation Models 13
    3.4.1 Beam Search 13
    3.4.2 Ensembling 14

**4 Fine-Tuning as a Domain Adaptation Baseline, and its Difficulties** 14
  4.1 Difficulty: Not Enough In-Domain Data 15
  4.2 Difficulty: Forgetting Previous Domains 15
  4.3 Difficulty: Overfitting to New Domains 16

**5 Data-Centric Adaptation Methods** 16
  5.1 Selecting Additional Natural Data for Adaptation 17
  5.2 Filtering Existing Natural Data 18
  5.3 Generating Synthetic Bilingual Adaptation Data from Monolingual Data 19
    5.3.1 Back Translation 19
    5.3.2 Forward Translation 20
  5.4 Artificially Noising and Simplifying Natural Data 21
  5.5 Synthetic Bilingual Adaptation Data 22

**6 Architecture-Centric Adaptation** 22
  6.1 Domain Labels and Control 23
  6.2 Domain-Specific Subnetworks 24
  6.3 Lightweight Added Parameters 25
  6.4 Architecture-Centric Multi-Domain Adaptation 26

**7 Training Schemes for Adaptation** 26
  7.1 Objective Function Regularization 27
    7.1.1 Freezing Parameters 27
    7.1.2 Regularizing Parameters 27
    7.1.3 Knowledge Distillation 29
  7.2 Curriculum Learning 29
  7.3 Instance Weighting 31
  7.4 Non-MLE Training 32
    7.4.1 Minimum Risk Training 32
    7.4.2 Meta-Learning 33

**8 Inference Schemes for Adaptation** 34
  8.1 Multi-Domain Ensembling 34
    8.1.1 Domain Adaptive Ensembling 34
    8.1.2 Retrieval-Based Ensembling 36
  8.2 Constrained Inference and Rescoring 36
  8.3 Domain Terminology Via Pre- and Post-Processing 38
    8.3.1 Terminology Tagging 38
    8.3.2 In-Domain Priming 38

**9 Case Studies for (Multi-)Domain Adaptation** 39
  9.1 Low Resource Language Translation as a Domain Adaptation Problem 39
    9.1.1 Can Low Resource Language Translation Be Treated as a Domain Adaptation Problem? 40
    9.1.2 Improving Low Resource Language Translation by Adaptation 40
  9.2 Gender Bias in Machine Translation as a Domain Adaptation Problem 41
    9.2.1 Can Gender Bias be Treated as a Domain Adaptation Problem? 42
    9.2.2 Mitigating the Effects of Gender Bias By Adaptation 42
  9.3 Document-Level Translation as a Domain Adaptation Problem 43
    9.3.1 Can Document NMT Be Treated as a Domain Adaptation Problem? 43
    9.3.2 Improving Document-Level Translation By Adaptation 44

**10 Conclusions and Future Directions** 45

**References** 46

## 1. Introduction

Neural Machine Translation (NMT) has seen impressive advances for some translation tasks in recent years. News and biomedical translation shared tasks from the Conference on Machine Translation (WMT) in 2019 already identified several systems as performing on par with a human translator for some high-resource language pairs according to human judgements (Barrault et al., 2019; Bawden et al., 2019). Indeed, these tasks involve not only high-resource language pairs but also relatively high-resource domains, with millions of relevant sentence pairs available for training. However, NMT models perform less well on

Saunders, D. (2022). *Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey*. arXiv. http://jair.org/index.php/jair/article/view/13566

# Domain/term adaptation in standard NMT and LLMs

**Customization options**

- ○ None
- ◒ TM
- ◓ Glossary
- ● Both

| Provider | Product | Customization |
|---|---|---|
| AI21 | Generative text AI | ● |
| Alibaba | eCommerce MT | ○ |
| Alibaba Cloud | General | ○ |
| Amazon | Translate | ● |
| Anthropic | Next-generation AI assistant | ● |
| Apptek | Neural Machine Translation | ○ |
| Baidu | Translate API | ◓ |
| Cloud Translation | Translation API | ○ |
| COTOHA (NTT) | Translator | ○ |
| DeepL | API | ◓ |
| Elia | Elhuyarren itzult-zaile automatikoa | ○ |
| Globalese | Machine Translation | ◒ |
| Google Cloud | Advanced Translation | ● |
| IBM Watson | eCommerce MT | ● |
| Kawamura by NICT | Translation Engine | ○ |
| Microsoft | Language Translator | ● |
| ModernMT | Static | ◒ |
| Mirai | Translator | ○ |
| Naver | Papago NMT Commercial | ○ |
| NiuTrans | Translation Cloud Platform | ○ |
| NVIDIA | NeMo framework | ◒ |
| OpenAI | GPT LLM x3 | ● |
| Oracle | Machine Translation | ○ |
| Pangeanic | Machine Translation API | ○ |
| PROMT | Cloud API | ○ |
| RoyalFlush Finance | Translation | ○ |
| SAP | Machine Translation | ○ |
| SYSTRAN | PNMT | ● |
| TartuNLP | Neurotõlge MT | ○ |
| Tencent Cloud | TMT API | ○ |
| Tilde | Machine Translation API | ○ |
| TREBE | Machine Translation API | ○ |
| Ubiqus | Translation API | ● |
| Yandex | Translate API | ● |
| Youdao | Cloud Translation API | ○ |

---

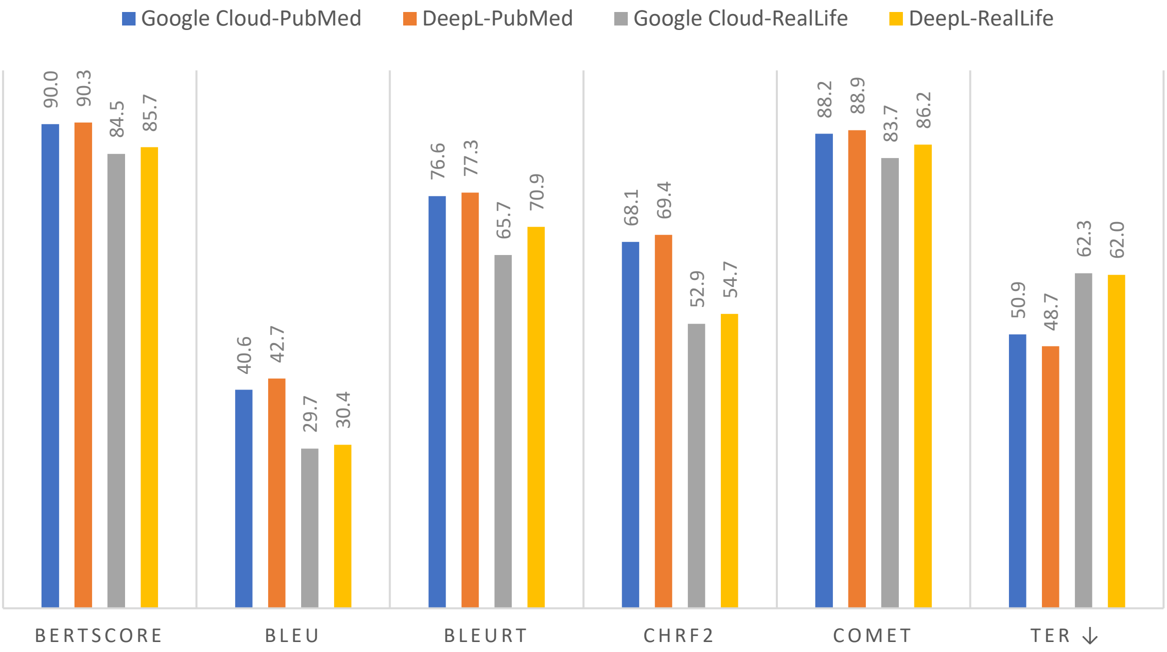
Large Language Models can be customized with TMs through fine-tuning, and terminology via prompt engineering

https://inten.to/machine-translation-report-2023

# Domain/term adaptation in LLMs: in-context learning

- Since domain/TB adaptation is implemented in traditional MT systems in various ways, and LLMs are capable of doing it in more than one way (e.g., by learning from examples or from term pairs, or by simply being told to treat the source text as medical), pairwise comparison may be the best approach in this setting.

# Baseline standard NMT: RealLife and PubMed



Vanroy, Bram and Tezcan, Arda and Macken, Lieve. (2023). MATEO: MAchine Translation Evaluation Online. In Nurminen, M. et al., *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* (pp. 499–500). https://lt3.ugent.be/mateo/

```
BLEU: nrefs:1|bs:1000|seed:12345|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1
chrF2: nrefs:1|bs:1000|seed:12345|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3
TER: nrefs:1|bs:1000|seed:12345|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version
BERTScore: nrefs:1|bs:1000|seed:12345|l:other|version:0.3.12|mateo:1.1.3
BLEURT: nrefs:1|bs:1000|seed:12345|c:BLEURT-20|version:commit cebe7e6|mateo:1.1.3
COMET: nrefs:1|bs:1000|seed:12345|c:Unbabel/wmt22-comet-da|version:2.0.1|mateo:1.1.3
```

| System | BERTScore | BLEU | BLEURT | chrF2 | COMET | TER ↓ |
|---|---|---|---|---|---|---|
| **Google Cloud-PubMed** | 90.0 | 40.6 | 76.6 | 68.1 | 88.2 | 50.9 |
| **DeepL-PubMed** | **90.3** | **42.7** | **77.3** | **69.4** | **88.9** | **48.7** |
| **Google Cloud-RealLife** | 84.5 | 29.7 | 65.7 | 52.9 | 83.7 | 62.3 |
| **DeepL-RealLife** | **85.7** | **30.4** | **70.9** | **54.7** | **86.2** | **62.0** |

# Baseline LLMs: sentence-by-sentence vs. batch prompting

Next, three LLMs (GPT4, Claude-2, and Bard) were asked to translate both sets (211 and 153 sentences) as single batches using the prompt schema initially introduced in Ghazvininejad, Gonen & Zettlemoyer (2023) and adopted in Peng et al. (2023):

```
You are a machine translation system.
Please provide the Russian translation for the following
sentences:
[English sentence 1]
...
[English sentence N]
```

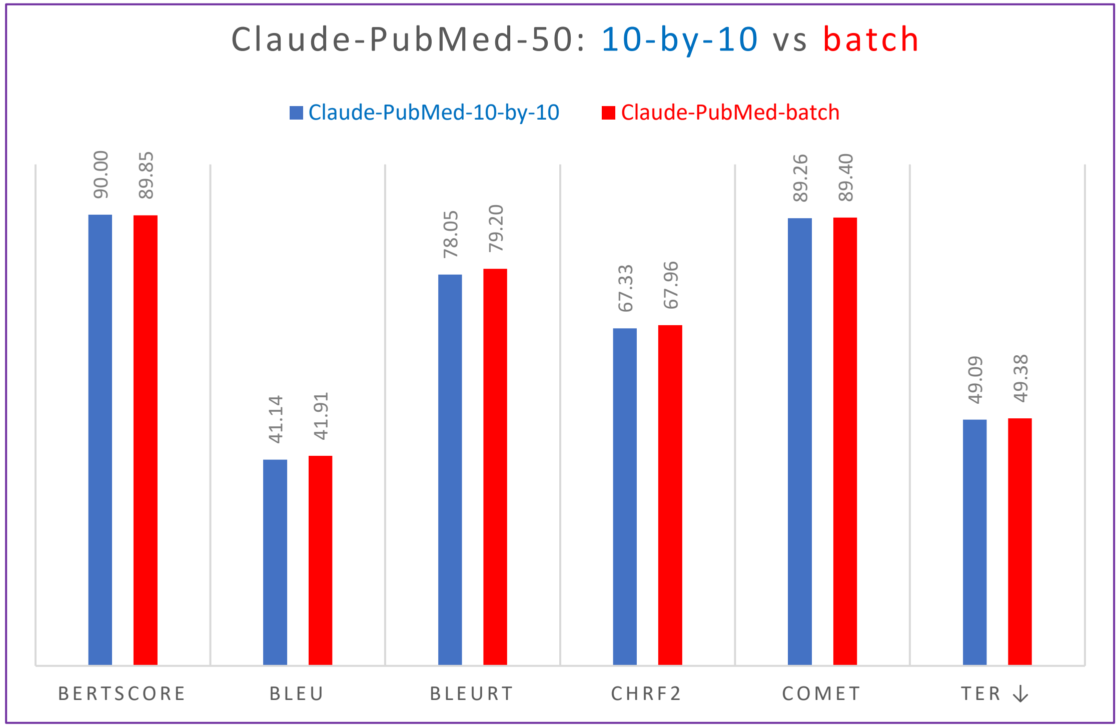
Positive reasons:

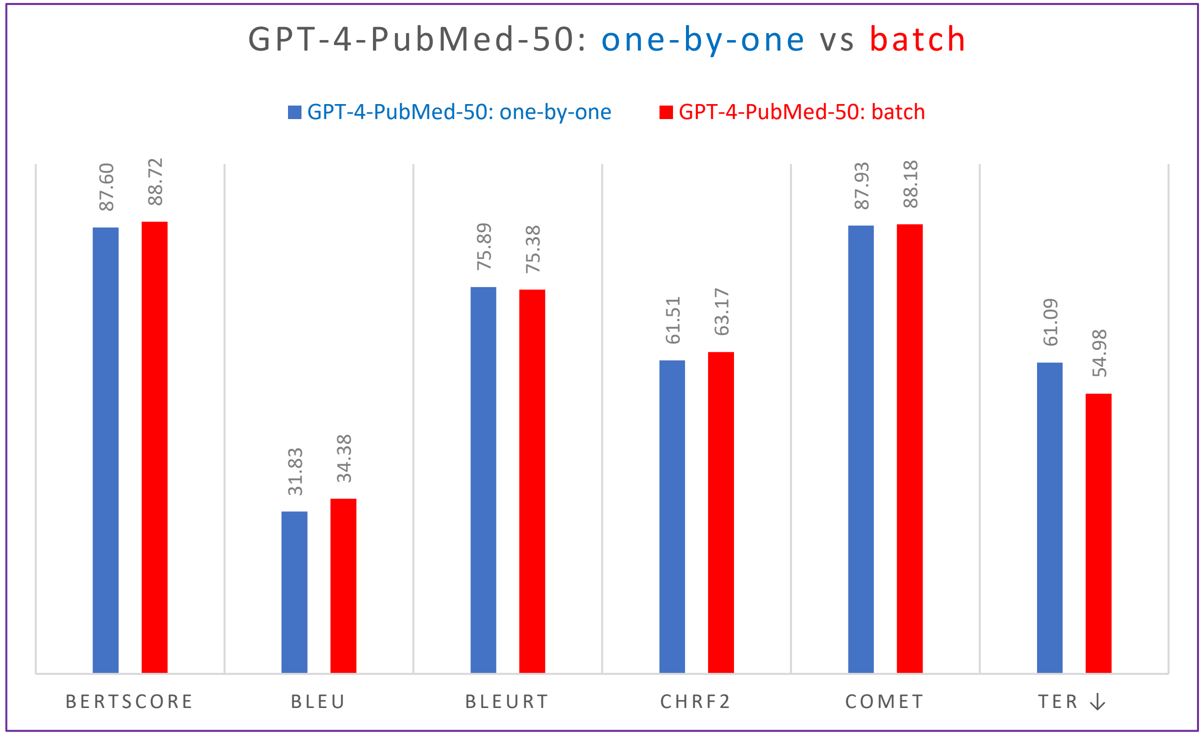
- User-friendly
- Time- (and cost-) efficient
- Batch prompting – interesting to explore
- May allow LLMs to learn more from context

Negative reason(s):

- Constraints associated with manual prompting (no API for Claude or Bard, etc.)

# Baseline LLMs: sentence-by-sentence vs. batch prompting





# Pushing LLMs to the limit…

GPT-4, Claude-2, and Bard were asked to translate both sets (211 and 153 sentences) as single batches using the prompt schema initially introduced in Ghazvininejad, Gonen & Zettlemoyer (2023) and adopted in Peng et al. (2023):

```
You are a machine
translation system.
Please provide the
Russian translation
for the following
sentences:
[English sentence 1]
...
[English sentence N]
```

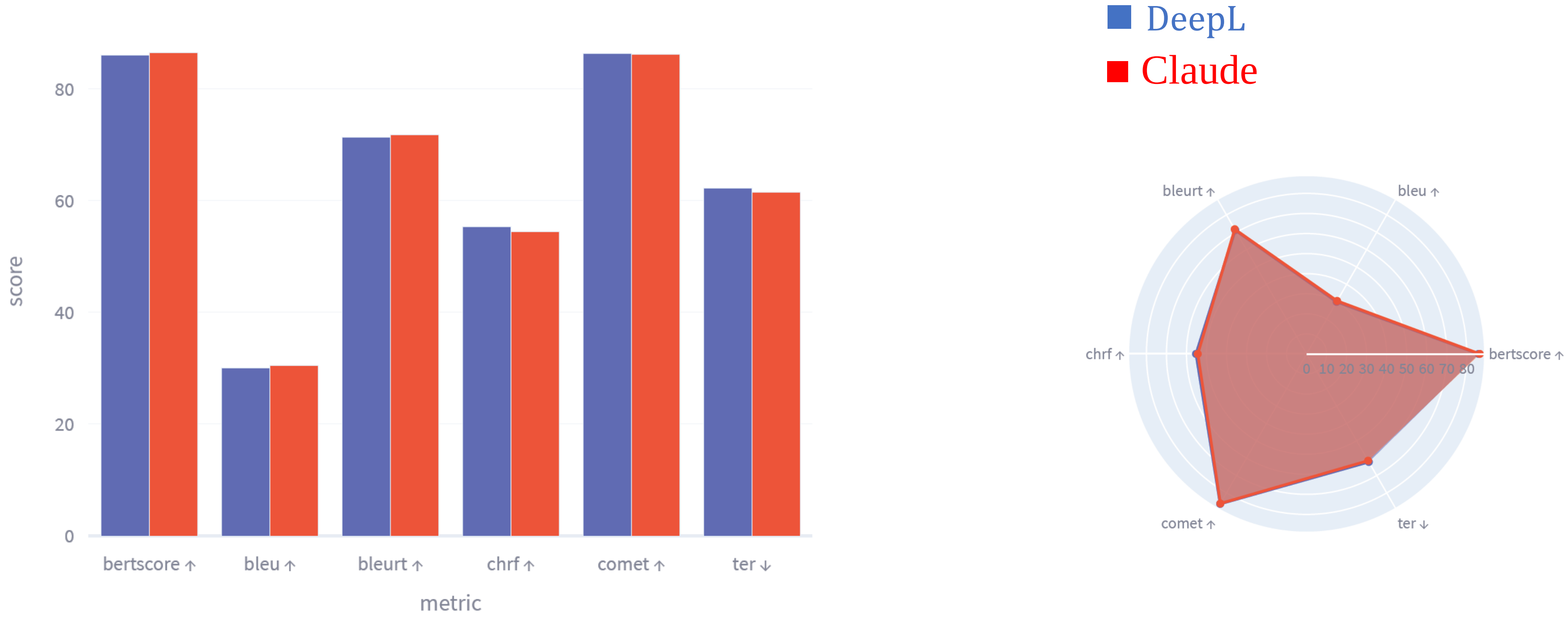
GPT-4:

“The provided text is quite long and detailed, containing numerous medical and trial-specific terms. This type of document would usually be translated by professional translators with expertise in clinical trials and medical terminology to ensure accuracy and adherence to regulatory standards…

“Given the specialized nature and length of the full text, I recommend utilizing professional translation services for the entire content to ensure accuracy, consistency, and compliance with medical translation standards…

”This is a lengthy and detailed text. Translating it in full while maintaining high accuracy and adhering to specific medical and trial terminology would require meticulous work, which is typically beyond the capabilities of an automated system, especially for such specialized content…”

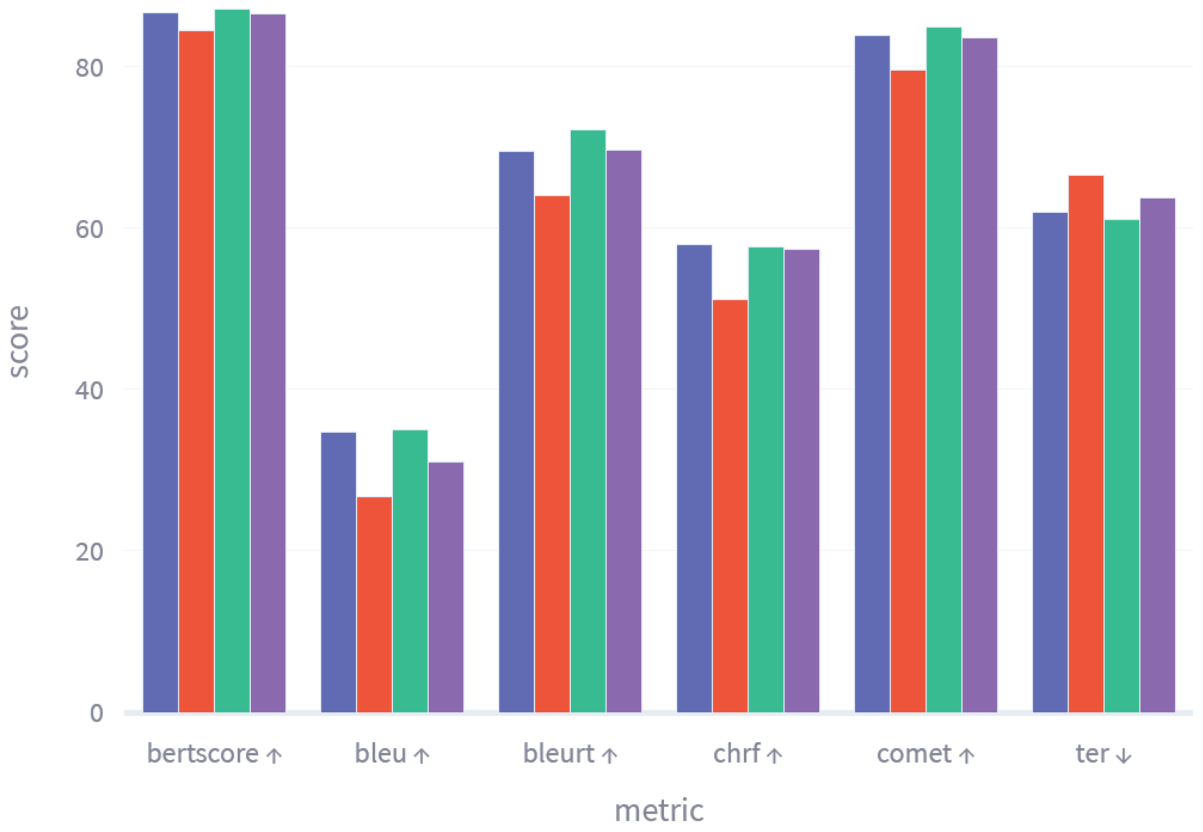
# DeepL and Claude: RealLife-131



| | BERTScore | BLEU | BLEURT | chrF2 | COMET | TER |
|---|---|---|---|---|---|---|
| **DeepL** | 85.94 | 29.99 | 71.25 | **55.23** | **86.33** | 62.19 |
| **Claude** | **86.50** | **30.51** | **71.78** | 54.35 | 86.13 | **61.47** |

DeepL and Claude. 131 EN-RU sentence pairs from RealLife. Single run with Claude. Evaluation with MATEO (Vanroy et al. 2023). Oct 13, 2023.

# DeepL, GPT-4, Claude, Bard. RealLife-50



🟦 DeepL
🟥 GPT-4
🟩 Claude
🟪 Bard

DeepL, GPT-4, Claude, Bard.
RealLife. 50 EN-RU sentence pairs.
Single runs for all LLMs. Evaluation
with MATEO (Vanroy et al. 2023).
Oct 12, 2023.

| | BERTScore | BLEU | BLEURT | chrF2 | COMET | TER |
|---|---|---|---|---|---|---|
| DeepL | 86.69 | 34.66 | 69.52 | **57.99** | 83.78 | 61.99 |
| GPT-4 | 84.48 | 26.69 | 63.95 | 51.07 | 79.52 | 66.52 |
| Claude | **87.17** | **35.04** | **72.14** | 57.58 | **84.85** | **61.12** |
| Bard | 86.58 | 30.97 | 69.57 | 57.38 | 83.61 | 63.71 |

# DeepL, GPT-4, Claude, Bard. RealLife-50: shuffled vs. unshuffled

| | BERTScore | BLEU | BLEURT | chrF2 | COMET | TER |
|---|---|---|---|---|---|---|
| | | | | | | |
| DeepL | 86.69 | **34.66** | 69.52 | 57.99 | **83.78** | **61.99** |
| DeepL-shuffled-1 | **87.05** | 33.97 | **69.72** | **58.09** | 83.57 | 62.85 |
| | | | | | | |
| GPT-4 | 84.48 | **26.69** | 63.95 | 51.07 | 79.52 | **66.52** |
| GPT-4-shuffled-1 | **85.44** | 26.16 | **63.98** | **51.60** | **81.12** | 68.25 |
| | | | | | | |
| Claude | **87.17** | **35.04** | **72.14** | **57.58** | **84.85** | **61.12** |
| Claude-shuffled-1 | 86.47 | 31.30 | 69.27 | 55.76 | 83.69 | 65.01 |
| | | | | | | |
| Bard | 86.58 | 30.97 | **69.57** | 57.38 | 83.61 | 63.71 |
| Bard-shuffled-1 | **87.60** | **37.78** | 69.40 | **59.95** | **83.84** | **60.91** |

# DeepL, GPT-4, Claude, Bard. RealLife-50: shuffled vs. unshuffled

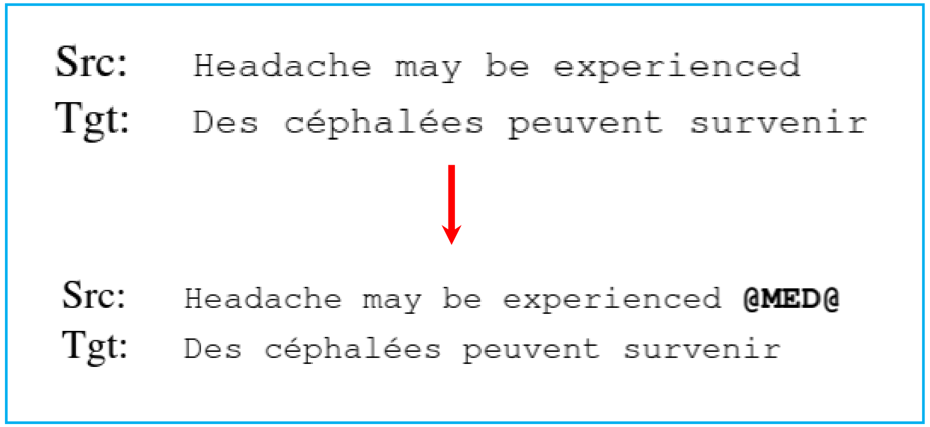
| | BERTScore | BLEU | BLEURT | chrF2 | COMET | TER |
|---|---|---|---|---|---|---|
| | | | | | | |
| DeepL | **86.69** | **34.66** | **69.52** | **57.99** | **83.78** | **61.99** |
| DeepL-shuffled-2 | 86.47 | 32.00 | 67.81 | 57.47 | 82.96 | 64.79 |
| | | | | | | |
| GPT-4 | 84.48 | 26.69 | 63.95 | 51.07 | 79.52 | **66.52** |
| GPT-4-shuffled-2 | 83.45 | **28.32** | **64.66** | **52.04** | **80.40** | 70.63 |
| | | | | | | |
| Claude | 87.17 | **35.04** | 72.14 | **57.58** | **84.85** | **61.12** |
| Claude-shuffled-2 | **87.32** | 32.12 | **72.81** | 57.11 | 83.89 | 65.87 |
| | | | | | | |
| Bard | 86.58 | 30.97 | 69.57 | **57.38** | **83.61** | 63.71 |
| Bard-shuffled-2 | **86.60** | **31.87** | **71.24** | 55.23 | 83.21 | **59.83** |

# Domain (name) adaptation in NMT and LLMs

## NMT

Can be done in many different ways (Saunders 2022):

- Fine-tuning the model with in-domain data
- Retraining the model from scratch on a mix of in- and out-of-domain data
- Enforcing desired terminology translation in pre- and/or post-processing (using statistical alignment?)
- Data augmentation:
  - In an early work, Kobus, Crego & Senellart (2017) proposed to implement domain control in NMT by adding additional tokens such as @MED@ to source sentences, or by combining word embedding with domain (name) embedding.



Src: `Headache may be experienced`
Tgt: `Des céphalées peuvent survenir`

Src: `Headache may be experienced` **`@MED@`**
Tgt: `Des céphalées peuvent survenir`

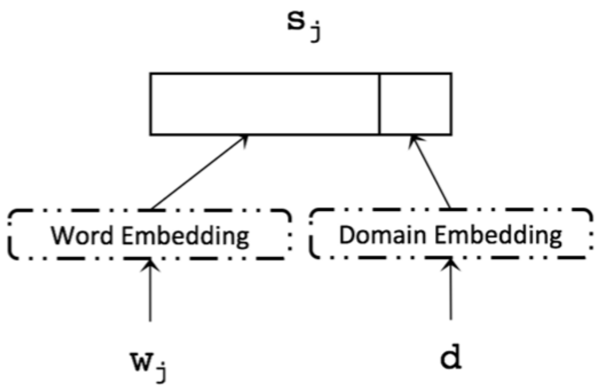


Figure 3: Word embedding layer for word $w_j$ extended with domain label $d$, which constitutes a new input $s_j$ for the encoder

# Domain (name) adaptation in NMT and LLMs

## NMT

Can be done in many different ways (Saunders 2022):

- Fine-tuning the model with in-domain data
- Retraining the model from scratch on a mix of in- and out-of-domain data
- Enforcing desired terminology translation in pre- and/or post-processing (using statistical alignment?)
- Data augmentation:
    - In an early work, Kobus, Crego & Senellart (2017) proposed to implement domain control in NMT by adding additional tokens such as @MED@ to source sentences, or by combining word embedding with domain (name) embedding.

## LLM

```
You are a machine translation
system that translates sentences
in the Clinical Trials domain.

    Please provide the Russian
    translation for the following
    sentences:
    [English sentence 1]
    ...
    [English sentence N]
```

# GPT-4, Claude, Bard: Medical vs. Baseline

| **PubMed-50** | BERTScore | BLEU | BLEURT | chrF2 | COMET | TER |
|---|---|---|---|---|---|---|
| | | | | | | |
| GPT-4 | **88.72** | **34.38** | **75.38** | **63.17** | **88.18** | **54.98** |
| GPT-4-Medical | 88.35<br>-0.4% | 33.68<br>-2.0% | 74.25<br>-1.5% | 61.70<br>-2.3% | 87.97<br>-0.2% | 56.51<br>-2.8% |
| | | | | | | |
| Claude | 90.00 | **41.14** | 78.05 | 67.33 | **89.26** | **49.09** |
| Claude-Medical | **90.11**<br>0.1% | 40.69<br>-1.1% | **78.35**<br>0.4% | **67.51**<br>0.3% | 89.25<br>0.0% | 49.16<br>-0.1% |
| | | | | | | |
| Bard | **88.38** | 35.76 | 75.26 | 63.85 | 87.38 | 55.27 |
| Bard-Medical | 88.27<br>-0.1% | **36.20**<br>1.2% | **75.98**<br>1.0% | **64.08**<br>0.4% | **87.65**<br>0.3% | **54.33**<br>1.7% |

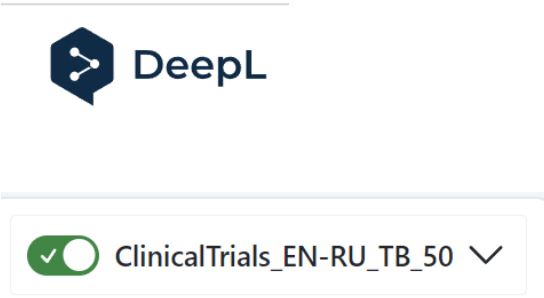
PubMed-50

RealLife-50

| **RealLife-50** | BERTScore | BLEU | BLEURT | chrF2 | COMET | TER |
|---|---|---|---|---|---|---|
| | | | | | | |
| GPT-4 | 84.48 | 26.69 | 63.95 | 51.07 | 79.52 | **66.52** |
| GPT-4-Medical | **86.47**<br>2.4% | **27.53**<br>3.1% | **68.63**<br>7.3% | **51.35**<br>0.5% | **83.12**<br>4.5% | 69.11<br>-3.9% |
| | | | | | | |
| Claude | 86.72 | **31.94** | **72.33** | **55.82** | **83.62** | **66.09** |
| Claude-Medical | **86.96**<br>0.3% | 31.57<br>-1.2% | 71.80<br>-0.7% | 55.32<br>-0.9% | 83.32<br>-0.4% | 66.31<br>-0.3% |
| | | | | | | |
| Bard | 86.58 | 30.97 | 69.57 | 57.38 | 83.61 | 63.71 |
| Bard-Medical | **89.01**<br>2.8% | **39.93**<br>28.9% | **76.29**<br>9.7% | **62.08**<br>8.2% | **85.95**<br>2.8% | **58.10**<br>8.8% |

# Terminology adaptation in NMT and LLMs

## NMT



DeepL

☑ ClinicalTrials_EN-RU_TB_50

## LLM

```
You are a machine translation system that translates
sentences in the Clinical Trials domain. In this
domain, the English terms below must be translated to
Russian as follows:

English: adverse events
Russian: нежелательные явления

English: cardiovascular events
Russian: сердечно-сосудистые события

...

English: [EN term N]
Russian: [RU Term N]

Using these requirements please translate the
following sentences to Russian:
[English sentence 1]
...
[English sentence K]
```

# Glossary vs. Baseline: PubMed-50

| **PubMed-50** | BERTScore | BLEU | BLEURT | chrF2 | COMET | TER |
|---|---|---|---|---|---|---|
| | | | | | | |
| DeepL-Baseline | 89.98 | 42.33 | 77.20 | 68.58 | 88.90 | 48.65 |
| DeepL-Glossary | **90.29** | **43.30** | **77.39** | **69.53** | **89.04** | **47.71** |
| | | | | | | |
| GPT-4-Baseline | **88.72** | 34.38 | **75.38** | **63.17** | **88.18** | **54.98** |
| GPT-4-Medical | 88.35 | 33.68 | 74.25 | 61.70 | 87.97 | 56.51 |
| GPT-4-Glossary | 88.62 | **34.48** | 74.90 | 63.11 | 87.73 | 55.13 |
| | | | | | | |
| Claude-Baseline | 90.00 | **41.14** | 78.05 | 67.33 | **89.26** | 49.09 |
| Claude-Medical | 90.11 | 40.69 | **78.35** | 67.51 | 89.25 | **49.16** |
| Claude-Glossary | **90.15** | 40.67 | 78.30 | **68.45** | 89.13 | 48.95 |

# Glossary vs. Baseline: RealLife-50

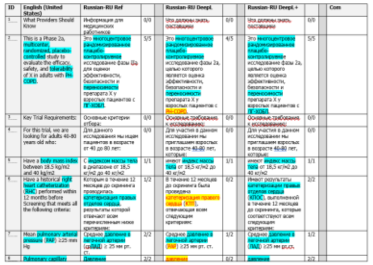
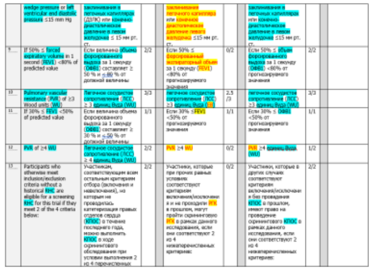
| **RealLife** | BERTScore | BLEU | BLEURT | chrF2 | COMET | TER |
|---|---|---|---|---|---|---|
| | | | | | | |
| DeepL | 86.69 | 34.66 | 69.52 | 57.99 | 83.78 | 61.99 |
| DeepL-Glossary | **90.96** | **42.63** | **82.09** | **64.59** | **89.28** | **54.00** |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| GPT-4 | 88.72 | 34.38 | 75.38 | 63.17 | 88.18 | 54.98 |
| GPT-4-Medical | 86.47 | 27.53 | 68.63 | 51.35 | 83.12 | 69.11 |
| GPT-4-Glossary | **92.33** | **42.41** | **82.70** | **63.38** | **89.39** | **52.48** |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| Claude | 90.00 | 41.14 | 78.05 | 67.33 | 89.26 | 49.09 |
| Claude-Medical | 86.96 | 31.57 | 71.80 | 55.32 | 83.32 | 66.31 |
| Claude-Glossary | **91.84** | **40.45** | **83.33** | **63.84** | **89.55** | **57.02** |
| | | | | | | |

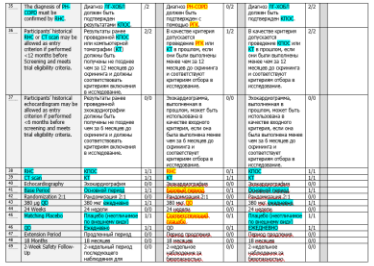
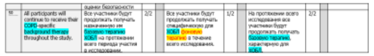
# LLM-Glossary vs. NMT-Glossary, RealLife-50: terminology recall

- Terminology Recall = the proportion of the occurrences of glossary terms translated exactly as required (and put in a correct grammatical form).
  - Some terms are multi-word, and some of the glossary translations are subjective. But the goal was to see how and to what extent the systems can handle them. E.g. ‘Base Period’ had to be translated as ‘Основной период’ not ‘Базовый период’.

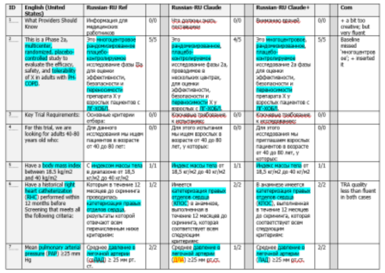
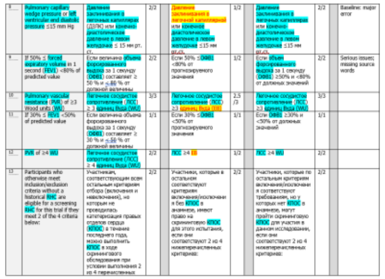
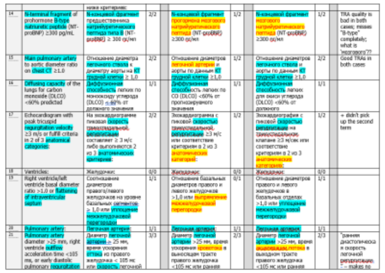
- Missing prepositions such as “давление заклинивания в легочных капиллярах” vs. “давление заклинивания легочного капилляра” was considered an error.
  - Similarly for “left ventricular end diastolic pressure”: “конечно-диастолическое давление в левом желудочке” vs. “конечное диастолическое давление левого желудочка”
- Where a correct non-abbreviated translation of a complex term was followed by the incorrect translation or failure to translate the acronym in parentheses, that was –0.5.
  - It was still interesting to see if giving the glossary to a system takes care of this.
- Incorrect word order in the translation of a complex term was considered an error:
  - E.g. translating ‘B-type natriuretic peptide’ as ‘B-типа натрийуретического пептида’ (correct: натрийуретического пептида типа B)

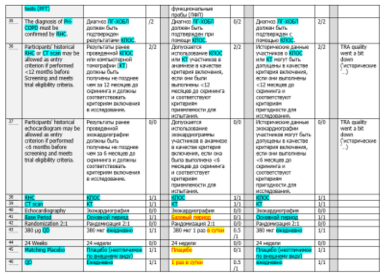
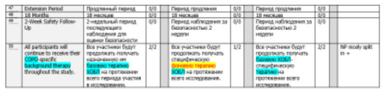
# LLM-Glossary vs. NMT-Glossary: RealLife-50 : terminology recall

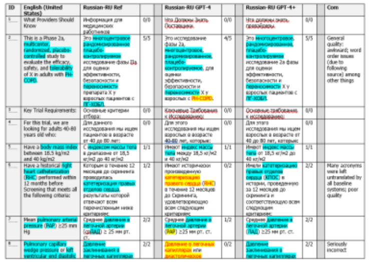
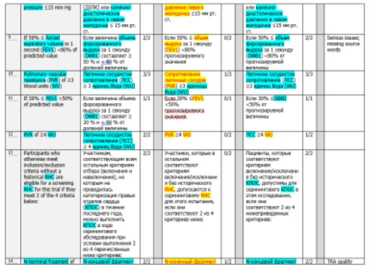
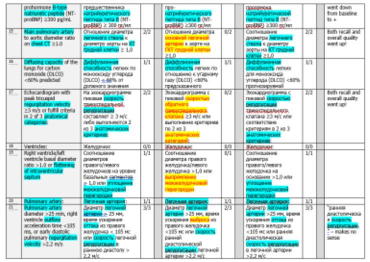
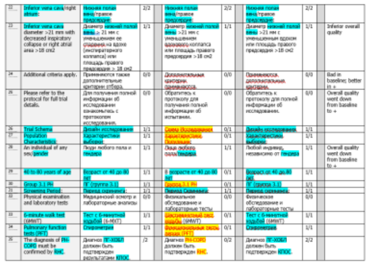
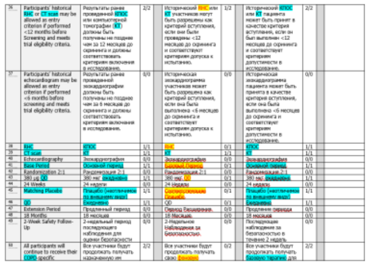



DeepL vs DeepL+

Claude vs Claude+

GPT-4 vs GPT-4+

# LLM-Glossary vs. NMT-Glossary, RealLife-50: terminology recall: some details

- Can enforcing correct terminology choices negatively impact the overall quality of TRA (both adequacy and fluency)?
- Handling of acronyms
- Reconciling term boundaries
- Proper NP splitting

**SRC**

Matching Placebo

**Ref**

Плацебо (неотличимое по внешнему виду)

**DeepL**

Соответствующий плацебо

**DeepL+**

Плацебо (неотличимое по внешнему виду)

**GPT-4**

Соответствующее Плацебо

**GPT-4+**

Плацебо (неотличимое по внешнему виду)

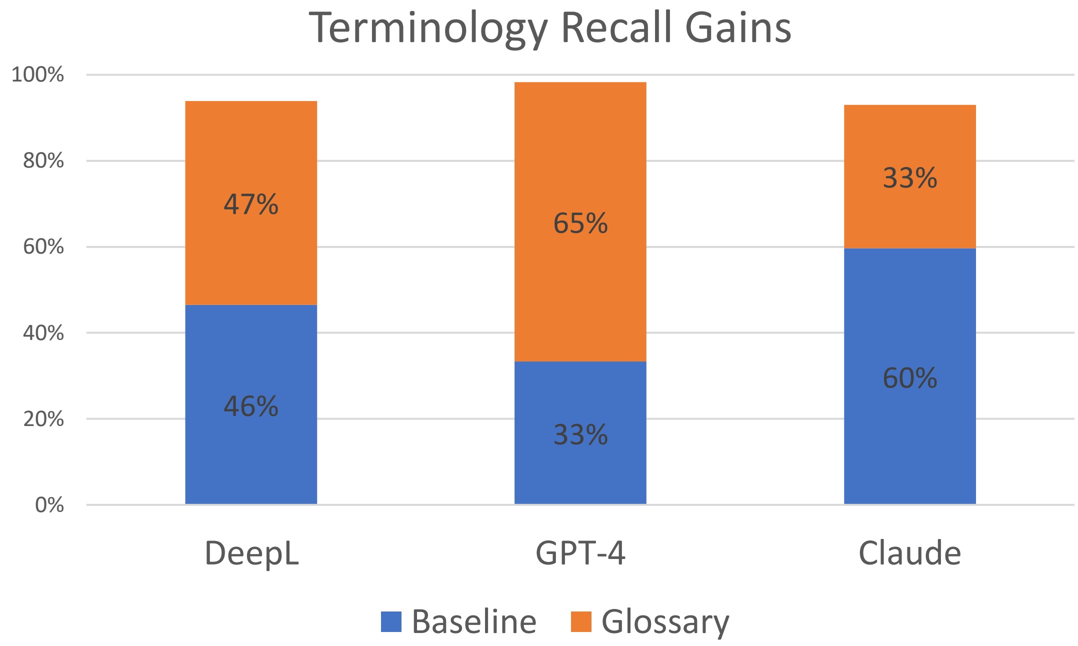
**Claude**

Плацебо

**Claude+**

Плацебо (неотличимое по внешнему виду)

# LLM-Glossary vs. NMT-Glossary: RealLife-50 : terminology recall



# Final thoughts

- Lots of ways to use LLMs in the human translation workflow, from making good use of TM matches to style change, shortening, grooming, adjusting the tone and register, data cleaning, and more.
- Lots of ways to use LLMs in MT post-editing, from asking them to perform MT QE to any amount of help with the post-editing process.
- My focus was narrow:
  - See how LLMs translate and how they adapt to domain and terminology constraints in a user framework, compared to standard MT systems.
  - Point fellow translators to amazing evaluation tools/resources they can use in their work.
- Further work:
  - Other language pairs and domains
  - Experiment with prompts and batch sizes
  - Again, let’s hope that we all will have API access to all these systems

# Some preliminary lessons

- Automatic scoring of a representative sample (even as short as 50 sentences) helps select the best LLM contender(s) for subsequent consideration.
- Comparing the scores for shuffled and unshuffled version may help select a systems that learn best from cross-sentential context.
- Batch prompting works in this setting and may help some LLMs to learn from context; no need to translate sentences one by one.
- Adding a domain label (such as ‘clinical trials’ or ‘pharmaceuticals’) to the prompt doesn’t help.
- All LLMs learn terminology from glossary-enriched prompts, with no systematic gain or loss to the overall TRA quality.
  - ➢But the overall quality of MT output is still far from perfect in all cases.
  - ➢Human expert in the loop is still sorely needed!

# Thanks!

- People:
  - My dear translation partner
  - Panos Kanavos, *NeuralDesktop*

- Entities:
  - Our respected client
  - MATEO: Vanroy et al. 2023
  - Franklin College of Arts & Sciences @ UGA
  - Philosophy and Cognitive Science of Deep Learning Group