

Compact Models for Machine Translation in the Bioinformatics Domain: Synthetic Data, Domain Adaptation, and Expert Evaluation

Jurgi Giraud^{1,2} and Andrew Gargett²

¹School of Informatics, University of Edinburgh, Edinburgh, UK,

²School of Languages and Applied Linguistics, The Open University, Milton Keynes, UK
jgiraud2@ed.ac.uk andrew.gargett@open.ac.uk

Abstract

Bioinformatics is almost invisible in Machine Translation (MT) research, despite being a hard case: dense specialised terminology, a high density of Multiword Expressions (MWEs), and an acute scarcity of parallel data. We address this gap for English–French. We compile a bioinformatics parallel corpus (14,467 sentence pairs) and expand it to 236k pairs through four augmentation strategies, each profiled for lexical diversity before use, then fine-tune five compact models (223M–8B parameters) under an incremental data ablation. We evaluate with automatic metrics and with a three-rater MQM evaluation involving professional life-sciences translators and using an error typology extended with a dedicated MWE category. Domain adaptation improves every system on every metric, with fine-tuned models surpassing much larger state-of-the-art commercial models on our domain test set. Expert judgements reveal a more structured picture with Accuracy errors falling by up to 68.8% and MWE errors by up to 51.1%, while also highlighting an accuracy–fluency trade-off that the automatic metrics conceal. Notably, the least lexically diverse augmentation sources contribute gains comparable to the most diverse ones, which we advance as one candidate account of the trade-off. We release all corpora, synthetic data, and adapted models.

1 Introduction

Machine Translation (MT) has become a routine part of how scientific knowledge circulates, but its coverage of specialised domains is uneven. Biomedical translation has been the subject of a sustained shared-task effort (Neves et al., 2018, 2024), while the neighbouring field of bioinformatics, now central to genomics, proteomics, and translational medicine, has received essentially no attention. At the outset of this work we could identify no bioinformatics parallel corpus, no domain-adapted MT

system, and no evaluation protocol designed for the domain’s characteristic linguistic profile.

This absence matters for three reasons. First, bioinformatics is where much of the life sciences’ methodological literature now lives, and its training and documentation materials are overwhelmingly English-first, a barrier for learners and practitioners working in other languages. Second, the domain is linguistically distinctive rather than simply “biomedical text with more code”: it combines dense specialised terminology with a high density of nominal MWEs (Giraud and Gargett, 2026), so that the units a translator must get right are frequently multi-token rather than single-token. Third, and most practically, parallel data in the domain is scarce enough that the standard recipe of “fine-tune on in-domain bitext” cannot be applied directly.

We also take a position on system scale. The prevailing assumption that translation quality tracks parameter count is a poor fit for institutional deployment in the life sciences, where compute budgets are modest, data governance may preclude commercial APIs, and MT is increasingly one component in a larger pipeline rather than an end-user product. Compact, specialised translators are attractive precisely because they can be fine-tuned cheaply, deployed locally, and orchestrated as tools inside agentic workflows, where the translation task is decomposed into drafting, checking, and editing stages handled by separate components (Briakou et al., 2024; Wu et al., 2024). Establishing whether a sub-4B domain-adapted model can match a frontier general-purpose system on a specialised domain is therefore a question about system design, not only about benchmark scores.

This paper reports a full pipeline for English–French bioinformatics MT and asks:

- **RQ1:** What impact do the selected domain-adaptation strategies, in particular synthetic data augmentation, have on the effectiveness

and efficiency of MT for bioinformatics?

- **RQ2:** To what extent does domain adaptation improve translation quality according to automatic metrics *and* to professional human judgement, with specific attention to terminology and MWEs?

Our contributions¹ are: (i) the first English–French bioinformatics parallel corpus and two comparable monolingual corpora, together with 174k synthetic pairs produced by four distinct strategies and profiled for lexical diversity before use; (ii) five domain-adapted systems spanning 223M–8B parameters, released openly, with a data ablation isolating the contribution of each augmentation source; (iii) an MQM-based human evaluation with three professional life-sciences translators, using an error typology extended with a dedicated MWE category and reporting high inter-rater reliability; and (iv) evidence that domain adaptation buys accuracy and terminological precision at a measurable cost in idiomaticity, a trade-off that the automatic metrics used here do not surface.

2 Background and Related Work

MT for specialised and life-science domains. Domain-specific MT has concentrated on “flag-ship” domains where large parallel corpora and institutional demand coincide: EU law (Koehn, 2005), patents, and broad scientific platforms such as SciELO (Soares et al., 2018). In the life sciences, the WMT biomedical task has driven most progress, working from Medline abstracts and clinical text (Neves et al., 2018; Bawden et al., 2019; Neves et al., 2023, 2024). Its recurring recipe is out-of-domain pre-training plus in-domain fine-tuning and data selection, and dedicated biomedical systems consistently beat generic ones on terminology and adequacy (Bérard et al., 2020; Liu and Huang, 2021). Two findings carry over to our setting: biomedical text has a long-tail terminological distribution that makes generic models suboptimal (Liu and Huang, 2021), and adapted systems are repeatedly observed to track source structure and lack the fluency of human translation (Neves et al., 2024), an observation our human evaluation quantifies. Bioinformatics shares the biomedical register but draws on computational, statistical, and software vocabulary alongside biological entities, limiting how far biomedical resources transfer.

¹An earlier version of this work appeared in the author’s doctoral thesis (Giraud, 2026).

Domain adaptation. General-purpose MT performs poorly on specialised text, with errors concentrated on exactly the lexical and phraseological phenomena specialists care about (Chu and Wang, 2018; Saunders, 2022). Fine-tuning on in-domain bitext is the dominant remedy (Sennrich et al., 2016), with evidence from legal MT (Bago et al., 2022), in-house translation-memory adaptation (Vieira et al., 2024), and scholarly communication, where Macken et al. (2024) found in a fine-grained human evaluation of English–French scientific translation that adapted systems made fewer terminology errors than a generic commercial baseline. Most relevant to our argument, Han et al. (2024) report that a small specialised clinical model outperformed a much larger general one *in expert judgement*. This is the pattern we reproduce for bioinformatics, and the reason we make expert evaluation rather than metrics the arbiter of the base-versus-adapted comparison. Fine-tuning nonetheless presupposes sufficient in-domain bitext and well-defined domain boundaries (Chu and Wang, 2018), neither of which holds here.

Data scarcity and augmentation. Low-resource conditions arise from under-resourced domains as well as language pairs (Haddow et al., 2022), and sparse in-domain data impedes the learning of domain terminology and MWEs (Haque et al., 2019). Back-translation of in-domain target text is a common response (Sennrich et al., 2016), effective across medical, legal, and IT domains (Kumari et al., 2021) and in biomedical shared-task submissions (Wu et al., 2022; Castaldo et al., 2024). LLMs now generate in-domain bitext directly under three broad strategies: paraphrasing, which rivals zero-shot generation (Radliński et al., 2025) but shows diminishing and eventually negative returns at volume (Sobrevilla Cabezudo et al., 2024); terminology-constrained generation, which nearly doubled correct term usage in Moslem et al. (2023); and free generation from scratch (Oh et al., 2023; Yang and Nicolai, 2023), scaled to build large synthetic corpora (de Gibert et al., 2025). Gains from added data saturate early, around 50k pairs in low-resource settings (Lee et al., 2022). Synthetic data also carries hallucinations and stylistic quirks that diverge from human translation norms (Nadăș et al., 2025). Lexical Diversity (LD) is one measurable proxy for this, and correlates with downstream quality in both LLM training (Chen et al., 2024) and back-translation for MT (Burchell et al., 2022);

we profile each of our synthetic sources on it in Section 3.4. What remains unestablished is how these strategies contribute *differentially*, since most studies test one against a natural-only baseline; our ablation separates them, and our human evaluation probes the stylistic cost they may carry.

Compact models and efficiency. Scaling laws favour larger models in general, but the heuristic that capability follows from scale is weakening (Hooker, 2025), and for task-constrained applications the evidence runs the other way. Wassie et al. (2024) show a domain-adapted 3.3B NLLB-200 beating fine-tuned 7–8B LLMs on biomedical translation, and Keles et al. (2024) report sub-100M biomedical models outperforming GPT-4-TURBO and Google Translate. Belcak et al. (2025) generalise this beyond MT, arguing that small models are the appropriate default for agentic workloads that repeat a narrow task under latency and cost constraints, a description that fits domain-specific MT, and that makes compact adapted translators natural tool-level components in orchestrated pipelines. Quantisation (Detmers et al., 2022) and parameter-efficient fine-tuning (Bapna and Firat, 2019; Hu et al., 2022) make this practical; QLoRA combines both, enabling single-GPU fine-tuning without degradation relative to 16-bit training (Detmers et al., 2023). In MT, Zhang et al. (2023) and Koneru et al. (2024) report consistent gains from PEFT, and, closest to our setup, Rios (2025) adapts FLAN-T5, LLAMA-3-8B, and TOWER-7B to the medical domain with 4-bit QLoRA. These choices are not only computational. Training a single large Transformer can emit hundreds of thousands of pounds of CO₂ (Strubell et al., 2019), and carbon accounting is now reportable (Lannelongue et al., 2021). Hence our restriction to sub-10B systems.

Terminology and MWEs. Specialised terminology is a documented MT failure point (Haque et al., 2020), and neural models are sensitive to domain shift partly because they lack parallel supervision for domain terms (Hu et al., 2019); evaluation of this is typically coarse, via term-list accuracy, and rarely covers multiword variants. Yet specialised concepts are frequently lexicalised as multiword units: multiword noun sequences are the predominant vehicle for technical term formation in English (Justeson and Katz, 1995), and scientific prose packs knowledge into dense nominal groups (Alves et al., 2024a). MWEs are the proverbial “pain in the neck” for NLP (Sag et al., 2002) because they

violate compositionality (Constant et al., 2017); in MT this surfaces as literal renderings and terminological inconsistency when systems fail to treat them as single units (Monti et al., 2018; Inurrieta et al., 2020), and linguistically motivated evaluations still report weaknesses on phraseology for systems competitive on aggregate scores (Manakhimova et al., 2023). For bioinformatics specifically, Giraud and Gargett (2026) show that nominal MWEs dominate the bioinformatics inventory, with distributions varying across IMRaD sections. Since English nominal compounds must generally be unpacked into prepositional or relative constructions in French, MWEs are a natural pressure point for EN–FR MT here, and one automatic metrics are poorly placed to detect, motivating our dedicated MQM category.

Evaluating domain-specific MT. Reference-based metrics remain standard (Papineni et al., 2002; Snover et al., 2006; Popović, 2015; Post, 2018), supplemented by neural metrics (Rei et al., 2020). Both have limits: reference dependence penalises valid alternatives, including renderings that follow domain convention more closely than the reference (Mathur et al., 2020), and neural metrics degrade under domain shift (Zouhar et al., 2024). More pointedly, general-purpose metrics give no fine-grained signal on specific phenomena, so terminological consistency and MWE rendering can improve substantially without this being visible in BLEU or COMET (Agrawal et al., 2024). Human evaluation therefore remains the reference point (Freitag et al., 2021), and error-typology annotation such as MQM (Lommel et al., 2013; Lommel, 2018; Freitag et al., 2021) yields a structured error profile supporting diagnosis rather than ranking alone, extensible to new domains through custom categories. Three recurring methodological concerns shaped our design: crowdsourced raters are contested for domain-specific assessment (Läubli et al., 2020; Krüger, 2022); reporting of rater profiles, instructions, and agreement is frequently incomplete (Marie et al., 2021); and sentence-level evaluation sits uneasily with the view that the unit of translation is the text in context (Krüger, 2022). We follow the first two (professional domain-specialist raters, transparent protocol, reliability statistics) and return to the third in Section 6.

3 Experimental Setup

3.1 Natural parallel data

No bioinformatics parallel corpus could be identified, so one was compiled. English–French abstract pairs were sourced from (i) research articles retrieved via the PubMed and Scopus APIs and scraped from publisher sites (EDP Sciences, ScienceDirect, John Libbey Eurotext), and (ii) francophone doctoral and master’s theses from HAL Thèses and theses.fr, which conventionally publish abstracts in both languages. This yielded 1,563 parallel abstracts (158 from articles, 1,405 from theses).

Abstracts were segmented with pySBD (Sadvikar and Neumann, 2020), aligned with Bertalign (Liu and Zhu, 2023), and filtered with Bicleaner AI (Zaragoza-Bernabeu et al., 2022) at a stringent threshold of 0.85 (against the 0.5 used by the tool’s authors) in order to retain high-quality pairs. The resulting corpus, BIOPAR, contains 14,467 sentence pairs, split 80/10/10 into train (11,567), development (1,450), and test (1,450).

3.2 Monolingual data

Two comparable monolingual corpora support augmentation and domain profiling. We utilised BIOMONO_EN² (Giraud and Gargett, 2026), a corpus comprising 4,707 bioinformatics articles from the alloplos collection (24.2M words), and BIOMONO_FR, comprising 212 French bioinformatics articles from ScienceDirect, *M/S*, and *Interstices* (343,834 words; 14,005 sentences).

3.3 Synthetic data generation

Four generation strategies were applied, chosen to target different weaknesses of the natural corpus, alongside a small quantity of mined bitext (Table 1).

Wikipedia bitext mining. EN/FR Wikipedia articles from six bioinformatics-adjacent categories were mined with LaBSE (Feng et al., 2022) sentence embeddings via the Sentence-Transformers mining framework (Reimers and Gurevych, 2020), retaining pairs above a similarity threshold of 1.65 (against the 1.2–1.3 usually recommended). 715 sentence pairs were retained. Unlike the four strategies below, this data is mined rather than generated, and is reported separately throughout.

Terminology-constrained generation. A bilingual terminology database was extracted with

²<https://github.com/jurgigi/BioMONO>

Source	Pairs
BIOPAR (train split), natural	11,567
Wikipedia bitext mining	715
Terminology-constrained generation	60,039
Paraphrasing (GPT-4o)	34,704
BIOMISTRAL over-generation	64,972
BIOMONO_FR back-translation	14,005
ELRC-EMEA (biomedical)	50,000
Total	236,002

Table 1: Composition of the training data.

tm2tb, a KeyBERT-based (Grootendorst, 2020) bilingual extractor, and used to constrain synthetic generation. Following Moslem et al. (2022, 2023), the top-1k single-, two-, and three-word terms from the bilingual terminology database were used to constrain generation: GPT-3.5-TURBO was prompted to produce 20 in-domain English sentences containing each source term, then to translate them under the constraint that the term be rendered by its approved French equivalent. 60,039 pairs.

Paraphrasing. GPT-4o generated three domain-constrained paraphrases per training-split English sentence, which were then translated. 34,704 pairs.

Open-model over-generation. To reduce reliance on commercial APIs, BIOMISTRAL-7B (Labrak et al., 2024) was used in an over-generation setup (Moslem et al., 2022): 15,000 randomly sampled French in-domain sentences from BIOMONO_FR were used as seeds (top- k = 10, top- p = 0.95, 2 returned sequences), producing roughly four synthetic sentences per seed. 64,972 pairs after back-translation.

Back-translation. BIOMONO_FR was back-translated with GPT-4o under a format-preserving prompt. 14,005 pairs were generated.

A further 50,000 pairs of neighbouring-domain biomedical data were drawn at random from ELRC-EMEA (Bicleaner-filtered at 0.85) to test whether adjacent-domain data still helps once in-domain synthetic data is abundant.

3.4 Lexical profiling of the synthetic data

Synthetic data can degrade a corpus while enlarging it: LLM outputs carry hallucinations, terminological slips, and stylistic quirks that diverge from human translation norms (Nadăș et al., 2025), and decoding strategies bias generation toward high-frequency phrasing (Holtzman et al., 2020). Lexical Diversity (LD) is one measurable dimension of this. Chen et al. (2024) report a positive correlation

Source	MTLD	HD-D
<i>Natural reference</i>		
BIOPAR_EN	89.81	0.8678
BIOMONO_EN	80.60	0.8898
<i>Synthetic</i>		
Back-translation	85.50	0.8537
Paraphrasing	65.24	0.8737
BIOMISTRAL over-generation	56.45	0.8371
Terminology-constrained	47.20	0.8490

Table 2: Lexical Diversity of each synthetic source against natural in-domain reference corpora.

between synthetic data diversity and downstream LLM performance, and in MT specifically Burchell et al. (2022) find that LD correlates more strongly with final model quality than syntactic diversity does. Work on recursive training shows the failure mode at the limit, with output diversity collapsing over successive generations (Guo et al., 2024; Shumailov et al., 2024; Zhang and Pavlick, 2025); our setting mixes synthetic with natural data and so is not that scenario, but the direction of risk is the same.

We therefore profiled each source before use with MTLD and HD-D, two measures designed to mitigate the length sensitivity of naïve type-token ratios (Malvern et al., 2004; McCarthy and Jarvis, 2010), computed with lexicalrichness. All scores in Table 2 are computed on the English side, so that synthetic sources and natural reference corpora are directly comparable; BIOPAR_EN denotes the English side of the natural parallel corpus. Wikipedia mining is excluded, as 715 pairs is too small for stable MTLD estimation.

Generation method drives LD far more than volume does (Table 2). Back-translation is the only source approaching natural text, exceeding the BIOMONO_EN reference on MTLD. Since these scores are computed on the generated English side, this cannot be attributed to natural prose: it reflects the seeding, as each sentence derives from a distinct naturally occurring French sentence and generation is constrained neither by a term list nor by a paraphrase target (Fadaee et al., 2017).

It is important to note that LD captures repetition, not adequacy; a highly diverse corpus can still be terminologically wrong. LD in generated text is also sensitive to prompt and decoding configuration (Martínez et al., 2024), so these figures characterise our generation setups rather than the strategies in the abstract. Section 5 returns to these figures once the ablation results are in view.

3.5 Ablation scenarios

Four incremental scenarios isolate the contribution of each data type³: **11k** (natural only); **107k** (+ Wikipedia, terminology-constrained, paraphrased); **186k** (+ BIOMISTRAL, back-translation); **236k** (+ ELRC-EMEA biomedical).

3.6 Models and fine-tuning

Three architectures were fine-tuned⁴ across all four scenarios: T5-BASE (223M) (Raffel et al., 2020), NLLB-200-DISTILLED-600M (NLLB Team et al., 2022), and LLAMA-3.1-8B-INSTRUCT (Touvron et al., 2023). Two further systems were fine-tuned on the 236k scenario only: MADLAD400-3B-MT (Kudugunta et al., 2023) and TOWERINSTRUCT-7B-v0.2 (Alves et al., 2024b). Encoder-decoder models were trained with Seq2SeqTrainer; decoder-only models with SFTTrainer and QLoRA (Dettmers et al., 2023). Full hyperparameters are in Appendix A.

3.7 Automatic evaluation

BLEU, chrF2, and TER are computed with SacreBLEU (Post, 2018), and COMET with WMT22-COMET-DA (Rei et al., 2020), on the held-out 1,450-pair in-domain test set. Comparison systems include non-adapted baselines, larger models in the same families, AYA-EXPANSE-8B, as well as commercial state-of-the-art and frontier models.⁵

3.8 Human evaluation

Thirty sentences ($\approx 1,000$ words) were sampled at random from BIOMONO_EN, within the 500–5,000-word range recommended for MQM linear scoring (Lommel et al., 2024). MWEs were identified by dependency parsing and list matching following Giraud and Gargett (2026), and marked inline with <mwe> tags so that raters could locate them; unsegmented raw source was also supplied for context.

The analysis reported here is restricted to base and domain-adapted pairs, so that the comparison isolates the effect of adaptation under matched architectures: T5-BASE vs. T5_236k, and LLAMA-3.1-8B-INSTRUCT vs. LLAMA-3.1-8B-INSTRUCT_236k. A fifth system, a commercial

³All training sets as well as development and test sets are available at <https://huggingface.co/jurgiraud/datasets>

⁴All fine-tuned models are available at <https://huggingface.co/jurgiraud/models>

⁵Commercial systems were accessed via their respective APIs.

baseline, was annotated in the same sessions but is excluded from this analysis, since it shares neither training data nor architecture with the adapted systems and would confound the base-versus-adapted contrast. Each rater therefore annotated 5,000 words in total. Three professional EN→FR translators specialising in the life sciences were recruited through the Société Française des Traducteurs, from a pool of 36 eligible translators identified across four professional bodies, and compensated at £0.05/word⁶ (£250 each), above prevailing post-editing and proofreading rates.

The MQM typology followed Freitag et al. (2021), with two severity levels (Minor, Major; multipliers 1 and 5) and main categories Terminology, Accuracy, Linguistic conventions, Style, Locale conventions, Source error, and Other. A dedicated **Multiword Expressions** category was added to capture errors on tagged MWEs. Annotation was carried out in a purpose-built Label Studio interface with a free-text comment field per error; comments were subsequently coded thematically (Braun and Clarke, 2022) in NVivo. Raw Quality Scores (RQS) were computed per system per rater, and inter-rater reliability assessed with two-way mixed-effects ICC.

4 Results

4.1 Data ablation

Table 3 shows a consistent picture across architectures. Every scenario improves on its baseline on every metric, and the largest single step is always the first: moving from no adaptation to 11k natural pairs gains T5-BASE 4.96 BLEU, NLLB 3.11, and LLAMA-3.1 1.88. For T5-BASE, this first step alone accounts for 79% of the total adaptation gain. Subsequent increments are far smaller and no longer follow a single pattern: for T5-BASE, 11k→107k adds 0.61 BLEU, 107k→186k adds 0.65, and 186k→236k adds 0.08; for NLLB the corresponding figures are 0.57, 0.04, and 0.05.

LLAMA-3.1 is the exception to the plateau: its largest post-baseline gain (+1.41 BLEU, −2.33 TER) comes at the final 186k→236k step, where the ELRC-EMEA biomedical data is added. This

⁶The amount provided to participants was based on a survey conducted by the Institute of Translation and Interpreting (ITI) Medical and Pharmaceutical Network (2022) which reports the various rates used by professional translators across translation-related tasks. Available at: https://itimmedical.co.uk/wp-content/uploads/2023/01/Translator_and_Interpreter_Survey_Results.pdf

System	BLEU↑	chrF2↑	TER↓	COMET↑
T5-BASE	39.23	67.31	51.27	84.50
+ 11k	44.19	71.15	45.98	85.43
+ 107k	44.80	71.57	46.24	85.60
+ 186k	45.45	71.99	45.62	85.61
+ 236k	45.53	71.92	45.62	85.64
NLLB-600M	41.00	68.48	49.26	84.54
+ 11k	44.11	70.90	46.72	85.41
+ 107k	44.68	71.33	46.42	85.74
+ 186k	44.72	71.53	46.38	85.75
+ 236k	44.77	71.64	46.24	85.84
LLAMA-3.1	40.39	69.41	50.36	85.01
+ 11k	42.27	69.68	48.83	85.30
+ 107k	42.69	69.83	48.71	85.30
+ 186k	43.16	70.46	48.01	85.27
+ 236k	44.57	71.44	45.68	85.55

Table 3: Data ablation across three architectures on the in-domain test set. Best per block in bold.

is consistent with a decoder-only instruction-tuned model benefiting from additional register-consistent data even after in-domain saturation, whereas the encoder–decoder systems have already converged.

Total adaptation gains are largest for the smallest model: T5-BASE gains 6.30 BLEU and 1.14 COMET, against 3.77/1.30 for NLLB and 4.18/0.54 for LLAMA-3.1. Baseline quality and headroom are inversely related, and architectural differences that are visible before adaptation largely disappear after it, with all three systems converging into a 44.5–45.5 BLEU band.

4.2 Cross-system comparison

Table 4 places the adapted systems against non-adapted open models and commercial baselines. The best system overall is MADLAD400-3B-MT_236k, a 3B model that outperforms every commercial system tested on all four metrics, including the frontier models GPT-5.5, Claude-SONNET-5, and Claude-OPUS-5. TOWER-7B_236k is close behind and posts the highest chrF2 of any system. Four of the five adapted systems, including the 223M T5-BASE, sit within 1.3 BLEU of GPT-4o, and all five outperform Google Translate by a wide margin.

Adaptation is also more effective than scale within families: T5_236k (223M) exceeds T5-LARGE (738M) by 5.08 BLEU, and NLLB-600M_236k exceeds NLLB-200-1.3B by 1.69. Notably, BIOMISTRAL-7B, pre-trained on biomedical text, is the weakest system tested (34.31 BLEU), a reminder that adjacent-domain pre-training is not a

System	BLEU $\uparrow$	chrF2 $\uparrow$	TER $\downarrow$	COMET $\uparrow$
<i>Domain-adapted (this work)</i>				
T5_236k	45.53	71.92	45.62	85.64
NLLB-600M_236k	44.77	71.64	46.24	85.84
LLAMA-3.1_236k	44.57	71.44	45.68	85.55
TOWER-7B_236k	46.18	72.83	45.32	86.29
MADLAD-3B_236k	47.44	72.72	43.78	86.38
<i>Non-adapted open systems</i>				
T5-LARGE	40.45	68.03	50.21	84.90
NLLB-200-1.3B	43.08	69.84	47.65	85.19
BiOMISTRAL-7B	34.31	64.59	56.47	82.41
MISTRAL-7B-INST	36.31	66.97	54.04	83.89
AYA-EXPANSE-8B	39.91	69.61	50.74	85.42
TOWER-7B-v0.2	42.65	70.39	48.46	85.73
MADLAD400-3B	43.64	71.01	46.44	86.15
<i>Commercial systems</i>				
Google Translate	40.29	69.31	49.03	83.30
GPT-3.5-TURBO	45.06	71.99	46.10	86.21
GPT-4-TURBO	45.56	72.34	45.61	86.29
GPT-4o	45.84	72.66	45.97	86.22
GPT-5.5	46.67	72.55	45.23	86.23
Claude-SONNET-5	46.87	72.05	45.41	85.97
Claude-OPUS-5	45.20	71.82	46.52	86.22

Table 4: All systems on the bioinformatics test set.

Category	T5			LLAMA-3.1		
	base	236k	Δ	base	236k	Δ
Accuracy	77	24	<i>−68.8%</i>	40	15	<i>−62.5%</i>
Terminology	31	21	<i>−32.3%</i>	52	27	<i>−48.1%</i>
MWes	45	22	<i>−51.1%</i>	34	26	<i>−23.5%</i>
Ling. conv.	18	20	<i>+11.1%</i>	54	16	<i>−70.4%</i>
Style	55	65	<i>+18.2%</i>	62	58	<i>−6.5%</i>
No error	12	30	<i>+150.0%</i>	10	26	<i>+160.0%</i>

Table 5: MQM error counts aggregated over three raters and 30 segments per system. Italics give the change from base to adapted.

substitute for task-specific adaptation, even though the same model proved useful as a synthetic data generator.

4.3 MQM evaluation

Table 5 aggregates error counts across the three raters. The pattern is sharply structured. Errors in the categories that matter most for technical adequacy fall steeply: Accuracy by 68.8% for T5 and 62.5% for LLAMA-3.1, Terminology by 32.3% and 48.1%, and MWes by 51.1% and 23.5%. Segments annotated error-free rise from 12 to 30 (T5) and from 10 to 26 (LLAMA-3.1).

Style behaves differently. It is the only category that does not improve for T5, rising by 18.2%, and it improves only marginally for LLAMA-3.1 (−6.5%). After adaptation, Style becomes the single largest error category for T5_236k (65 of the

	T5			LLAMA-3.1		
	base	236k	Δ	base	236k	Δ
Rater 1	64.46	78.62	<i>+21.97%</i>	67.52	84.52	<i>+25.18%</i>
Rater 2	87.37	94.91	<i>+8.63%</i>	84.22	95.72	<i>+13.65%</i>
Rater 3	74.64	87.78	<i>+17.60%</i>	66.19	86.56	<i>+30.78%</i>
Mean	75.49	87.10	<i>+16.07%</i>	72.64	88.93	<i>+23.20%</i>
≥ 1 Major	74.44	38.89	<i>−47.8%</i>	68.89	35.56	<i>−48.4%</i>

Table 6: MQM Raw Quality Scores per rater, and percentage of segments containing at least one Major error (mean over raters). For the RQS rows, Δ in the **Mean** row is the mean of the three per-rater percentage changes; for the Major-error row, Δ is computed on the rater means.

152 errors distributed across the five main categories), having been roughly comparable to Accuracy before. Linguistic conventions diverge between architectures: LLAMA-3.1 improves dramatically (−70.4%, from a very high base of 54), while T5 worsens slightly.

Table 6 reports quality scores. Domain adaptation raises RQS by 16.07% for T5 and 23.20% for LLAMA-3.1 on average, with every rater showing the effect in the same direction. The proportion of segments containing at least one Major error roughly halves for both architectures: 74.44% $\rightarrow$ 38.89% for T5 and 68.89% $\rightarrow$ 35.56% for LLAMA-3.1. The two adapted systems converge on a similar quality level (87.10 and 88.93) despite a 36-fold difference in parameter count, mirroring the convergence already visible in the automatic ablation.

Inter-rater reliability is high: ICC(3,1) = 0.85 (95% CI 0.47–0.98, $p < 0.01$) and ICC(3,k) = 0.95 (95% CI 0.72–0.99, $p < 0.01$), in the excellent range on standard interpretations (Koo and Li, 2016). Absolute severity differs between raters, with Rater 1 consistently the strictest, but relative system rankings are stable across all three.

4.4 Thematic analysis of translator comments

Coding of the free-text comments produced six top-level themes: Lexical Accuracy and Appropriateness; Syntactic and Grammatical Errors; Omission and Addition; Stylistic and Idiomatic Quality; Morphological and Orthographic Issues; and Multiword Expressions.

The lexical theme is the densest, with mistranslated adjectives, adverbs, and verbs alone accounting for 81 references across all four systems. Comments on base-system output typically concern out-

right failures of meaning; comments on adapted output concern finer distinctions of register and precision, for example flagging *grande* where *très grande* or *immense* would render *huge*. The distinction is diagnostic: adaptation moves the failure mode from misunderstanding to imprecision.

The MWE theme confirms the quantitative pattern while adding nuance. Base systems omit or mistranslate MWEs in ways that distort meaning, as when *in competition with each other* is rendered *en compétition* with the reciprocal element dropped. Adapted systems generally preserve MWE boundaries and components but sometimes select French realisations that raters mark as non-idiomatic rather than incorrect.

The stylistic theme is dominated by the code “grammatically correct but overliteral or unnatural,” together with specific recurring complaints, notably the overuse of the preposition *pour* in constructions where French would prefer a different linker. Raters also flagged inconsistent terminology within a single system’s output, a document-level failure that segment-level training does not address.

5 Discussion

Bigger doesn’t necessarily mean better. On automatic metrics, compact models below 4B parameters match and at times outperform much larger open and commercial models, including frontier models such as GPT-5.5 and Claude-OPUS-5. This carries practical implications for life-sciences institutions, as compact domain-adapted models can be served locally on a single consumer GPU while keeping potentially sensitive data inside the institution, and at a fraction of the training footprint of the largest system we fine-tuned (Table 7). Compact models are also more tractable components for agentic and multi-agent pipelines, where translation is one call among several and per-call latency and cost compound, thus directly contributing to current conversations regarding compact models and their practical advantages as agents (Belcak et al., 2025). The corollary is that the relevant investment is data, not parameters. Our results suggest that the first 10k in-domain pairs deliver most of the achievable gain, that roughly 100k pairs bring systems close to their ceiling, and that further scaling of either data or model size returns little.

Different synthetic sources do different work. The ablation lets us attribute effects, though the pattern differs by architecture. For NLLB,

the 11k→107k step, driven by terminology-constrained generation and paraphrasing, is the largest post-baseline gain; for T5-BASE it is essentially tied with the following step; and for LLAMA-3.1 the largest gain comes last. Where the 11k→107k step does help, it improves chrF2 and TER alongside BLEU, which suggests genuine structural learning rather than n-gram inflation. The 107k→186k step adds back-translated in-domain French and yields smaller but consistent gains, most plausibly on target-side fluency. The final 236k step adds neighbouring-domain biomedical data and helps least for the encoder–decoder systems, consistent with the core bioinformatics phenomena already being covered, while helping LLAMA-3.1 most. Terminology-constrained generation and back-translation are thus complementary rather than interchangeable: the former densifies coverage of the terms the domain actually uses, the latter supplies naturally occurring target-side phrasing.

The lexical profile in Section 3.4 complicates this picture. The two sources dominating the 107k scenario are by some margin the least lexically diverse of the four (MTLD 47.20 and 65.24, against 80.60–89.81 for natural in-domain text), yet they deliver gains comparable to the far more diverse back-translated data added at 186k. This sits awkwardly with the positive diversity–performance association reported by Chen et al. (2024) and Burchell et al. (2022). The likely reconciliation is that terminology-constrained data trades diversity for term coverage, which is what a terminology-dense domain most needs. Whether that trade is worthwhile depends on which errors one is willing to tolerate, a question the automatic metrics cannot answer and the human evaluation can.

Terminology and MWEs are where domain adaptation pays off. The dedicated MWE category earns its place in the typology. MWE errors fall by half for T5 and by a quarter for LLAMA-3.1, and the qualitative data shows the mechanism: adapted systems learn to keep multi-token units intact and to select their conventional French equivalents, where base systems decompose or drop them. Given that nominal compounds dominate the domain’s MWE inventory and generally require restructuring in French, this is precisely the phenomenon a bioinformatics-specific system needs to handle, and precisely the one that a BLEU difference of a few points fails to make visible.

An accuracy–fluency trade-off that metrics conceal. The most consequential finding is the divergence between what the metrics and the raters see. On automatic metrics, T5_236k improves monotonically. On human judgement, its accuracy improvements are accompanied by an 18.2% *increase* in Style errors, so that stylistic quality becomes its dominant residual weakness. Adaptation appears to pull the output toward the literal, term-faithful end of the spectrum, and in doing so toward *translationese*. This matches the accuracy–fluency trade-off formalised by Lim et al. (2024) and the information-theoretic account of Flamich et al. (2025), and aligns with the biomedical MQM profile reported by Zouhar et al. (2024), where fluency errors substantially outweigh accuracy errors, in contrast to news-domain benchmarks.

Our synthetic data is a plausible contributor, and Section 3.4 gives the suspicion some empirical footing. Back-translated source text carries the artefacts of the back-translating system, and the two sources dominating the 107k scenario score well below natural in-domain text on LD, indicating training material that is more repetitive and formulaic than the prose the systems are ultimately asked to produce. Fine-tuning on such pairs may teach terminology and structure while simultaneously teaching a slightly off register. We stress that this is an association rather than a demonstrated cause: our four sources differ on many dimensions besides diversity, and LD in generated text is itself sensitive to prompt and decoding configuration (Martínez et al., 2024). If the account holds, the trade-off is not intrinsic to adaptation but a property of how the adaptation data was built, and could be attacked by filtering synthetic data on stylistic as well as semantic criteria, or by a dedicated fluency-oriented editing stage downstream of translation.

Implications for evaluation. On automatic metrics, adaptation looks uniformly positive: every scenario improves every score for every architecture. The raters see something more differentiated, and locate a regression in a category that the metrics do not isolate. Any evaluation of domain-specific MT that reports only sentence-level metrics will therefore overstate how close adapted systems are to being deployment-ready. The practical recommendation that follows from the human study is that domain adaptation raises the floor of quality substantially, but does not remove the need for stylistic post-editing where publication-quality French is

required.

6 Limitations

Single language pair. All experiments are EN→FR. French is a well-resourced target with an established scientific register and existing terminological infrastructure; the same pipeline applied to a lower-resourced or morphologically richer target would face different constraints, particularly in terminology-constrained generation, where the assumption of a single approved target form is weaker. Extension to additional targets is the most direct next step.

Synthetic data quality is only partially assessed. We profile each synthetic source for Lexical Diversity (Section 3.4) and measure the downstream effect of each on translation quality, but we do not assess the intrinsic adequacy, terminological correctness, or register fidelity of the synthetic pairs themselves. LD captures repetition, not whether a generated sentence is well-formed, accurate, or stylistically typical of the domain. Given our evidence that the least diverse sources deliver gains comparable to the most diverse, a fuller pre-use audit, combining LD with adequacy checks, register comparison against natural in-domain text, and hallucination detection, would allow the trade-off to be attributed with confidence rather than inferred.

Human evaluation scale. Three raters, 30 segments, and four systems is at the lower bound of the range recommended by Freitag et al. (2021), chosen under budget constraints. ICC values are high but their confidence intervals are wide (ICC(3,1): 0.47–0.98), and absolute severity varies noticeably between raters. Additionally, the two best systems on automated metrics, MADLAD400-3B_236k and TOWER-7B_236k, were not included in the human study. A larger study with more raters, more segments (and therefore longer context), and full system coverage would allow significance testing rather than descriptive comparison.

7 Conclusion

We have presented the first end-to-end study of machine translation for bioinformatics, covering corpus construction, four synthetic augmentation strategies profiled for lexical diversity, adaptation of five compact and mid-sized systems, and expert evaluation under an MWE-extended MQM protocol. Domain adaptation works, and works

best where it matters: a 3B adapted model outperforms every commercial system we tested, including current frontier models, on the domain test set, and adapted systems reduce Accuracy errors by roughly two-thirds and MWE errors by a quarter to a half in professional judgement. It also has a cost that automatic metrics do not report: stylistic quality does not improve alongside accuracy and can degrade. For a domain where terminological precision is non-negotiable but the output will often be read by specialists rather than published, this is a favourable trade, but it is one that should be made knowingly, and it points toward adaptation pipelines that treat fluency as a separately optimised objective rather than a by-product. All corpora, synthetic datasets, and adapted models are released openly.

Ethics Statement

The human evaluation received institutional ethics approval (reference number 2024-0259-2). Participants gave informed consent, were free to withdraw, and were compensated for their time. Annotations are reported in aggregate and by anonymised rater ID.

Sustainability Statement

The carbon footprint of the fine-tuning runs was estimated using the *Green Algorithms* framework proposed by Lannelongue et al. (2021). Across all runs, fine-tuning consumed 137.5 kWh for an estimated 31.8 kgCO₂e; the per-model breakdown is given in Table 7. LLAMA-3.1-8B-INSTRUCT alone accounts for 56% of the total, while MADLAD400-3B, the best-performing system, accounts for 6.5%. These figures exclude synthetic data generation, which was performed on external API infrastructure and could not be measured.

Model	kWh	kgCO ₂ e
T5-BASE	19.72	4.56
NLLB-200-DISTILLED-600M	17.21	3.98
LLAMA-3.1-8B-INSTRUCT	76.57	17.70
TOWERINSTRUCT-7B-v0.2	15.00	3.47
MADLAD400-3B	8.97	2.07
Total	137.47	31.78

Table 7: Estimated energy use and carbon footprint of all fine-tuning runs.

References

- Sweta Agrawal, António Farinhas, Ricardo Rei, and André Martins. 2024. [Can automatic metrics assess high-quality translations?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502.
- Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024a. [Multi-word expressions in English scientific writing](#). In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 67–76.
- Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024b. [Tower: An open multilingual large language model for translation-related tasks](#). *Preprint*, arXiv:2402.17733.
- Petra Bago, Sheila Castilho, Edoardo Celeste, Jane Dunne, Federico Gaspari, and Andy Way. 2022. [Sharing high-quality language resources in the legal domain to develop neural machine translation for under-resourced European languages](#). *Revista de Llengua i Dret, Journal of Language and Law*, (78):9–34.
- Ankur Bapna and Orhan Firat. 2019. [Simple, scalable adaptation for neural machine translation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548.
- Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurélie Névél, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. [Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53.
- Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. [Small language models are the future of agentic AI](#). *Preprint*, arXiv:2506.02153.
- Alexandre Bérard, Zae Myung Kim, Vassilina Nikoulina, Eunjeong L. Park, and Matthias Gallé. 2020. [A multilingual neural machine translation model for biomedical data](#). In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.
- Virginia Braun and Victoria Clarke. 2022. *Thematic Analysis: A Practical Guide to Understanding and Doing*. SAGE Publications.

- Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. [Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317.
- Laurie Burchell, Alexandra Birch, and Kenneth Heafield. 2022. [Exploring diversity in back translation for low-resource machine translation](#). In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 67–79.
- Antonio Castaldo, Maria Zafar, Prashanth Nayak, Rejwanul Haque, Andy Way, and Johanna Monti. 2024. [The SETU-ADAPT submission for WMT 24 biomedical shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 647–653.
- Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abidin. 2024. [On the diversity of synthetic data and its impact on training large language models](#). *Preprint*, arXiv:2410.15226.
- Chenhui Chu and Rui Wang. 2018. [A survey of domain adaptation for neural machine translation](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319.
- Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. [Multiword expression processing: A survey](#). *Computational Linguistics*, 43(4):837–892.
- Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. [Scaling low-resource MT via synthetic data generation with LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27674–27692.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. [LLM.int8\(\): 8-bit matrix multiplication for transformers at scale](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.
- Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. [Data augmentation for low-resource neural machine translation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.
- Gergely Flamich, David Vilar, Jan-Thorsten Peter, and Markus Freitag. 2025. [You cannot feed two birds with one score: The accuracy-naturalness tradeoff in translation](#). *Preprint*, arXiv:2503.24013.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Jurgi Giraud. 2026. [Developing Machine Translation for Bioinformatics: An Exploration into Domain-Specific Terminology, Domain Adaptation, and Evaluation](#). PhD thesis, The Open University.
- Jurgi Giraud and Andrew Gargett. 2026. [Beyond single words: MWE identification in bioinformatics research articles and dispersion profiling across IMRAD](#). In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 86–95. Association for Computational Linguistics.
- Maarten Grootendorst. 2020. [KeyBERT: Minimal keyword extraction with BERT](#).
- Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. [The curious decline of linguistic diversity: Training language models on synthetic text](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604.
- Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. [Survey of low-resource machine translation](#). *Computational Linguistics*, 48(3):673–732.
- Lifeng Han, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic. 2024. [Neural machine translation of clinical text: An empirical investigation into multilingual pre-trained language models and transfer-learning](#). *Frontiers in Digital Health*, 6.
- Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. 2019. [Terminology translation in low-resource scenarios](#). *Information*, 10(9):273.
- Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. 2020. [Analysing terminology translation errors in statistical and neural machine translation](#). *Machine Translation*, 34(2):149–195.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). *Preprint*, arXiv:1904.09751.
- Sara Hooker. 2025. [On the slow death of scaling](#). SSRN Scholarly Paper No. 5877662.

- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. [Domain adaptation of neural machine translation by lexicon induction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001.
- Uxoá Inurrieta, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka, and Kepa Sarasola. 2020. [Learning about phraseology from corpora: A linguistically motivated approach for multiword expression identification](#). *PLOS ONE*, 15(8):e0237767.
- John S. Justeson and Slava M. Katz. 1995. [Technical terminology: Some linguistic properties and an algorithm for identification in text](#). *Natural Language Engineering*, 1(1):9–27.
- Bunyamin Keles, Murat Gunay, and Serdar I. Caglar. 2024. [LLMs-in-the-loop part-1: Expert small AI models for bio-medical text translation](#). *Preprint*, arXiv:2407.12126.
- Philipp Koehn. 2005. [Europarl: A parallel corpus for statistical machine translation](#). In *Proceedings of Machine Translation Summit X*, pages 79–86.
- Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. [Contextual refinement of translations: Large language models for sentence and document-level post-editing](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2711–2725.
- Terry K. Koo and Mae Y. Li. 2016. [A guideline of selecting and reporting intraclass correlation coefficients for reliability research](#). *Journal of Chiropractic Medicine*, 15(2):155–163.
- Ralph Krüger. 2022. [Some translation studies informed suggestions for further balancing methodologies for machine translation quality evaluation](#). *Translation Spaces*, 11(2):213–233.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. [MADLAD-400: A multilingual and document-level large audited dataset](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.
- Surabhi Kumari, Nikhil Jaiswal, Mayur Patidar, Manasi Patwardhan, Shirish Karande, Puneet Agarwal, and Lovekesh Vig. 2021. [Domain adaptation for NMT via filtered iterative back-translation](#). In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 263–271.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. [BioMistral: A collection of open-source pretrained large language models for medical domains](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864.
- Loïc Lanelongue, Jason Grealey, and Michael Inouye. 2021. [Green algorithms: Quantifying the carbon footprint of computation](#). *Advanced Science*, 8(12):2100707.
- Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. [A set of recommendations for assessing human-machine parity in language translation](#). *Journal of Artificial Intelligence Research*, 67:653–672.
- En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David I. Adelani, Ruisi Su, and Arya D. McCarthy. 2022. [Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation?](#) In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67.
- Zheng Wei Lim, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. 2024. [Simpson’s paradox and the accuracy-fluency tradeoff in translation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–103.
- Boxiang Liu and Liang Huang. 2021. [ParaMed: A parallel corpus for English–Chinese translation in the biomedical domain](#). *BMC Medical Informatics and Decision Making*, 21(1):258.
- Lei Liu and Min Zhu. 2023. [Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts](#). *Digital Scholarship in the Humanities*, 38(2):621–634.
- Arle Lommel. 2018. [Metrics for translation quality assessment: A case for standardising error typologies](#). In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, pages 109–127. Springer International Publishing.
- Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. [The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control](#). In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas*, pages 75–94.
- Arle R. Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. [Multidimensional quality metrics: A flexible system for assessing translation quality](#). In *Proceedings of Translating and the Computer 35*.

- Lieve Macken, Vanessa De Wilde, and Arda Tezcan. 2024. [Machine translation for open scholarly communication: Examining the relationship between translation quality and reading effort](#). *Information*, 15(8).
- David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan.
- Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. [Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT?](#) In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245.
- Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. [Scientific credibility of machine translation research: A meta-evaluation of 769 papers](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306.
- Gonzalo Martínez, José Alberto Hernández, Javier Conde, Pedro Reviriego, and Elena Merino-Gómez. 2024. [Beware of words: Evaluating the lexical diversity of conversational LLMs using ChatGPT as case study](#). *ACM Transactions on Intelligent Systems and Technology*.
- Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. [Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997.
- Philip M. McCarthy and Scott Jarvis. 2010. [MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment](#). *Behavior Research Methods*, 42(2):381–392.
- Johanna Monti, Violeta Seretan, Gloria Corpas Pastor, and Ruslan Mitkov, editors. 2018. *Multiword Units in Machine Translation and Translation Technology*, volume 341. John Benjamins Publishing Company.
- Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. [Domain-specific text generation for machine translation](#). In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30. Association for Machine Translation in the Americas.
- Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. [Domain terminology integration into machine translation: Leveraging large language models](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911. Association for Computational Linguistics.
- Mihai Nadăș, Laura Dioșan, and Andreea Tomescu. 2025. [Synthetic data generation using large language models: Advances in text and code](#). *IEEE Access*, 13:134615–134633.
- Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéal, Stefan Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. [Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138.
- Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéal, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. [Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54.
- Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéal, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. [Findings of the WMT 2018 biomedical translation shared task: Evaluation on Medline test sets](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 1 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Seokjin Oh, Su Ah Lee, and Woohwan Jung. 2023. [Data augmentation for neural machine translation using generative language model](#). *Preprint*, arXiv:2307.16833.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Maja Popović. 2015. [chrF: Character n-gram f-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.
- Łukasz Radliński, Mateusz Guściora, and Jan Kocoń. 2025. [Backtranslation and paraphrasing in the LLM](#)

- era? comparing data augmentation methods for emotion classification. In *Computational Science – ICCS 2025*, pages 3–17. Springer Nature Switzerland.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.
- Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.
- Miguel Rios. 2025. Instruction-tuned large language models for machine translation in the medical domain. In *Proceedings of Machine Translation Summit XX, Volume 1: Research Track*, pages 162–172. European Association for Machine Translation.
- Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of the Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114.
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer Berlin Heidelberg.
- Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.
- Ilya Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. The curse of recursion: Training on generated data makes models forget. *Preprint*, arXiv:2305.17493.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.
- Felipe Soares, Viviane Moreira, and Karin Becker. 2018. A large parallel corpus of full-text scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association.
- Marco Antonio Sobrevilla Cabezudo, Marcio Lima Inacio, and Thiago A. S. Pardo. 2024. Investigating paraphrase generation as a data augmentation strategy for low-resource AMR-to-text generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 663–675.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.
- Inacio Vieira, Will Allred, Seamus Lankford, Sheila Castilho, and Andy Way. 2024. How much data is enough data? fine-tuning large language models for in-house translation: Performance evaluation across multiple dataset sizes. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 236–249.
- Aman Kassahun Wassie, Mahdi Molaei, and Yasmin Moslem. 2024. Domain-specific translation with open-source large language models: Resource-oriented analysis. *Preprint*, arXiv:2412.05862.
- Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141.
- Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, Yuanchang Luo, Yuhao Xie, Miaomiao Ma, Ting Zhu, Lizhi Lei, Song Peng, Hao Yang, and Ying Qin. 2022. HW-TSC translation systems for the WMT22 biomedical translation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 936–942.
- Wayne Yang and Garrett Nicolai. 2023. Neural machine translation data generation and augmentation using ChatGPT. *Preprint*, arXiv:2307.05779.
- Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831. European Language Resources Association.

Lingjun Zhang and Ellie Pavlick. 2025. [Does training on synthetic data make models less robust?](#) In *Proceedings of the Sixth Workshop on Insights from Negative Results in NLP*, pages 79–85.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. [Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jinyuan Wang, and Brian Thompson. 2024. [Fine-tuned machine translation metrics struggle in unseen domains](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500.

A Fine-tuning parameters

A.1 Training setups

A.1.1 T5-BASE and NLLB-200-DISTILLED-600M (Seq2SeqTrainer)

Both T5-BASE and NLLB-200-DISTILLED-600M were fine-tuned with the transformers Seq2SeqTrainer using the same training configuration:

- **Epochs:** 8
- **Learning rate:** 2×10^{-5}
- **Batch size:** 16 (per device)
- **Precision:** bfloat16 (bf16)

A.1.2 LLAMA-3.1-8B-INSTRUCT (SFTTrainer + QLoRA)

LLAMA-3.1-8B-INSTRUCT was fine-tuned using transformers SFTTrainer with QLoRA (4-bit quantization + LoRA adapters):

LoRA configuration

- **Rank (r):** 8
- **α :** 16
- **Dropout:** 0.05
- **Target modules:** ["q_proj", "v_proj", "o_proj"]

Training configuration

- **Epochs:** 4
- **Learning rate:** 1×10^{-4}
- **Batch size:** 16 (per device)
- **Gradient accumulation:** 2
- **Optimizer:** paged_adamw_8bit

Chat template

```
<|im_start|>user
{USER_PROMPT}<|im_end|>
<|im_start|>assistant
{MODEL_RESPONSE}<|im_end|>
```

A.2 Total runtime across all runs

Model	Total runtime (hh:mm:ss)
T5-BASE	33:31:12
NLLB-200-DISTILLED-600M	29:03:18
LLAMA-3.1-8B-INSTRUCT	307:44:12

Table 8: Total fine-tuning runtime aggregated across all runs for each model.

A.3 Compute environment

A.3.1 T5-BASE and NLLB-200-DISTILLED-600M

- **Platform:** Google Colab⁷
- **CPU:** Intel(R) Xeon(R)
- **GPU:** NVIDIA A100-SXM4-40GB
- **RAM:** 83.5 GB

A.3.2 LLAMA-3.1-8B-INSTRUCT

- **Operating system:** Windows 11 (Desktop)
- **CPU:** Intel Core i7-14700F
- **GPU:** NVIDIA GeForce RTX 4060 Ti
- **RAM:** 64 GB (DDR4)

⁷<https://colab.research.google.com/>