Meta4BR: Avaliando a Fidelidade Metafórica em Traduções de Metáforas para o Português por LLMs

Luisa Stellet, Isabella Leite, Gabriel Assis, Aline Paes

Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ, Brasil
{luisastellet@id,isabellalps@id,assisgabriel@id,alinepaes@ic}.uff.br

Abstract. *Metaphors are key elements of human communication, enabling nuanced expression and deeper conceptual understanding. Yet, most research on metaphor detection focuses on English corpora and models. The ability of Large Language Models (LLMs) to process metaphorical language in other languages remains underexplored. This study investigates whether metaphors can be preserved and annotated through translation. Focusing on Brazilian Portuguese, we propose a pipeline based on back-translation and comparison using multiple LLMs. Based on both automatic metrics and human evaluation, results suggest that conceptual and linguistic metaphors can be effectively translated. The findings suggest that metaphor resources can be bootstrapped via translation workflows. Nonetheless, challenges remain, particularly the need for extensive manual validation and the risks of cultural bias and semantic drift.*

Resumo. *Metáforas são elementos fundamentais na comunicação humana, permitindo uma expressão nuançada e uma compreensão conceitual mais profunda. No entanto, a maioria das pesquisas em detecção de metáforas concentra-se em corpora e modelos para o inglês. A capacidade dos Grandes Modelos de Língua (LLMs) em processar linguagem metafórica em outros idiomas permanece pouco explorada. Este estudo investiga se metáforas podem ser preservadas e anotadas por meio da tradução. Com foco no português brasileiro, adotamos um pipeline baseado em retrotradução e comparação, utilizando múltiplos LLMs. Com base em métricas automáticas e avaliação humana, resultados sugerem a viabilidade de traduzir metáforas conceituais e linguísticas. Os achados apontam que recursos de metáforas podem ser construídos a partir de fluxos de tradução. No entanto, permanecem desafios, especialmente a necessidade de validação manual extensiva e os riscos de viés cultural e mudança semântica.*

1. Introdução

Metáforas são figuras de linguagem em que termos são usados fora de seu sentido literal para expressar uma relação implícita de semelhança entre conceitos distintos [Glucksberg and Keysar 1993]. Longe de se limitar ao campo literário, a metáfora é central na comunicação humana, facilitando a compreensão de ideias abstratas e refletindo aspectos do pensamento conceitual [Lakoff and Johnson 2008]. Assim, espera-se que um sistema projetado para comunicação, incluindo modelos de língua neurais (LMs) [Bengio et al. 2003, Vaswani et al. 2017, Brown et al. 2020], seja capaz de processar e empregar metáforas de modo eficaz [Goatly 1997].

Entretanto, a avaliação sistemática da capacidade de LMs lidarem com metáforas depende de recursos linguísticos específicos para esse fenômeno. Tal demanda é par-

ticularmente desafiadora em línguas com menos recursos disponíveis, como o português brasileiro, dado o forte componente cultural e linguístico das figuras de linguagem [Kövecses 2010]. Essa falta de recursos abertamente disponíveis limita a avaliação de habilidades complexas de LMs, dificultando a sua adoção em contextos culturais específicos [Joshi et al. 2020]. Porém, a construção de um *corpus* dedicado de metáforas exige considerável investimento em tempo e especialização [Shutova 2017]. Nesse cenário, aproveitar recursos já disponíveis em inglês para gerar um *corpus* de metáforas em português brasileiro surge como uma alternativa promissora, embora envolta em desafios linguísticos, culturais e metodológicos [Xia et al. 2019, Ranathunga et al. 2023].

Por outro lado, metáforas, enquanto fenômenos linguísticos e cognitivos, representam um desafio particular na tradução [Liu et al. 2025]. A partir disso, focamos em abordar metáforas conceituais e metáforas linguísticas. A metáfora conceitual é uma forma de compreender um conceito abstrato, presente na estrutura mental dos falantes. Na metáfora "`tempo é dinheiro`", o conceito abstrato de "`tempo`" é entendido a partir do mais concreto "`dinheiro`", por meio de comparação implícita. Nela, distinguem-se dois domínios: o domínio-fonte, derivado da experiência humana e responsável por transferir sentido, e o domínio-alvo, mais abstrato e receptor desse sentido. Nesse caso, "`dinheiro`" é o domínio-fonte e "`tempo`" o domínio-alvo, pois o primeiro atribui valor ao segundo. Assim, a metáfora define-se pelo mapeamento entre domínios.

Por sua vez, as metáforas linguísticas correspondem a suas manifestações em expressões concretas [Deignan 2016]. A frase "`estou economizando tempo`" revela, no discurso, a visão do tempo como um recurso que pode ser economizado, assim como o dinheiro. Dessa forma, é necessário considerar o contexto em que as metáforas são utilizadas e os discursos em que estão inseridas. Embora algumas metáforas conceituais tenham caráter potencialmente universal, como associações entre "`para cima`" e positividade, suas realizações linguísticas variam significativamente entre culturas, refletindo especificidades histórico-sociais [Kövecses 2005]. Um exemplo ilustrativo é "*`She shoot down her arguments`*" *vs.* "`Ela destruiu seus argumentos`". Embora o domínio-fonte da metáfora seja o mesmo (guerra), o tipo de ação usado difere (atirar/derrubar *vs.* destruir), o que poderia trazer uma distorção semântica se traduzido literalmente.

Este artigo propõe uma abordagem para avaliar a viabilidade de construir um *corpus* de metáforas em português a partir de recursos em inglês. A proposta combina dois eixos de validação: horizontal e vertical. Horizontalmente, avaliamos a concordância entre traduções (inglês $\rightarrow$ português) de diferentes modelos, considerando convergências como indicadores de validade linguística e divergências como potenciais sinais de ambiguidade. Verticalmente, aplicamos retrotradução (do inglês *backtranslation*; inglês $\rightarrow$ português $\rightarrow$ inglês), para verificar a fidelidade das traduções, observando se o conteúdo original é preservado por meio da comparação entre o texto original e sua retrotradução.

Avaliamos nossa abordagem em duas bases: uma de metáforas conceituais e outra de metáforas linguísticas. A análise inclui seis LLMs de propósito geral, majoritariamente de menor porte por razões de custo computacional e preocupações ambientais [Morrison et al. 2025], além de três modelos ajustados para tradução. Os resultados são medidos por sete métricas de similaridade léxica, semântica e de qualidade de tradução, complementadas por análise humana detalhada [Freitag et al. 2021, Wang et al. 2024] e por uma medida estatística de concordância.

Os valores obtidos nas métricas de retrotradução, a análise de concordância entre os diferentes modelos e a avaliação humana apontam que há potencial na criação e validação de bases de metáforas em português a partir de recursos em inglês[1].

2. Trabalhos Relacionados

Esta seção aborda trabalhos relacionados voltados à tradução automática de metáforas e à construção de recursos linguísticos que envolvem linguagem figurada.

2.1. Traduções de Metáforas

A tradução automática de metáforas é um tema interdisciplinar que envolve linguística, estudos da tradução e processamento de linguagem natural (PLN). Sua natureza não literal, culturalmente situada e ambígua impõe desafios adicionais à tradução computacional [Postigo 2024, Wang et al. 2024]. Estudos anteriores abordam desde traduções interlinguísticas específicas [Labarta Postigo 2023, Ma 2025] até comparações entre tradutores humanos [Massey 2021, Labarta Postigo 2023], sistemas neurais de tradução automática (*Neural Machine Translation systems*, NMTs) tradicionais [Massey 2021, Labarta Postigo 2023, Ma 2025] e LLMs [Postigo 2024, Wang et al. 2024].

[Massey 2021] mostrou que NMTs tendem a simplificar ou omitir metáforas conceituais entre inglês e alemão, ao contrário de tradutores humanos, que produzem versões mais expressivas. Para o português, [Labarta Postigo 2023] apontou falhas do DeepL e Systran na tradução de metáforas esportivas, apesar da fluência lexical. No contexto técnico, [Ma 2025] propôs um NMT otimizado para metáforas em textos sobre *Big Data*, destacando desafios conceituais persistentes.

Mais recentemente, LLMs têm apresentado avanços. [Postigo 2024] relatou melhor desempenho estilístico e retórico do ChatGPT em comparação a NMTs na tradução para espanhol e português, embora metáforas culturais permaneçam problemáticas. No estudo MMTE [Wang et al. 2024], um *corpus* anotado, avaliou-se a tradução figurativa entre inglês, chinês e italiano, evidenciando ganhos semânticos com LLMs como GPT-4o e Gemini-1.0-pro, mas com limitações culturais, especialmente no chinês.

Este trabalho amplia essas investigações ao explorar, em escala, a capacidade de LLMs na tradução metafórica entre inglês e português. Consideramos uma diversidade de modelos, com diferentes *corpora* de pré-treinamento, incorporando múltiplas perspectivas linguísticas. Além disso, avaliamos modelos abertos e de menor porte, investigando seu potencial como alternativas mais acessíveis e inclusivas.

2.2. Criação de Bases de Dados de Metáforas

A construção de *datasets* para identificação e análise de metáforas impõe desafios metodológicos relevantes [Boisson et al. 2023]. Ferramentas automáticas exigem cautela para mitigar vieses e garantir qualidade [Boisson et al. 2023]. A escassez de recursos compartilhados e públicos limita a difusão de *corpora* metafóricos, ainda poucos em volume e diversidade frente a outras áreas do PLN [Boisson et al. 2025]. Entre os disponíveis, destaca-se o *VU Amsterdam Metaphor Corpus* (VUAMC) [Krennmayr and Steen 2017], baseado no método MIPVU [Steen 2010], amplamente adotado em estudos posteriores [Boisson et al. 2025, Boisson et al. 2023].

[1] *Scripts* e dados estão disponíveis em https://github.com/MeLLL-UFF/meta4br-mt/

Apesar de diretrizes como o MIPVU, a anotação manual permanece custosa [Zayed et al. 2020, Joseph et al. 2023]. Exemplos incluem o *corpus* chinês de 30 mil sentenças de [Shao et al. 2024] e o trabalho de [Tong et al. 2024], que utiliza até cinco anotadores via *crowdsourcing* para avaliar o desempenho de LLMs em inglês. Estratégias híbridas também têm sido exploradas: [Zayed et al. 2020] combinam recuperação automática e medidas semânticas, enquanto [Joseph et al. 2023] empregam LLMs para gerar rótulos iniciais (*silver labels*) a serem refinados manualmente.

Esforços recentes ampliam o foco para línguas sub-representadas. [Kabra et al. 2023] desenvolvem um *corpus* com sete idiomas de baixo recurso, e [Sanchez-Bayona and Agerri 2024] propõem o Meta4XNLI para inferência textual com metáforas em inglês e espanhol, combinando traduções humanas e anotações automáticas. Para o português, [Boldarine 2024] produz um *corpus* bilíngue com aprendizes escolares. Mais próximo à abordagem deste estudo, o MMTE [Wang et al. 2024] inclui inglês, italiano e chinês, gerando traduções metafóricas via NMTs e GPT-4o.

Avançando nessa direção, exploramos o uso de LLMs comerciais e abertos de menor porte, para investigar seu potencial na criação de recursos metafóricos em português. Buscamos ampliar a disponibilidade de dados anotados, conciliando diversidade, eficiência e qualidade, com suporte de métricas automáticas e validação humana.

3. Meta4BR: Método para Avaliar Traduções de Metáforas para o Português

Este artigo investiga a habilidade de modelos de língua neurais, incluindo LLMs de propósito geral e modelos especializados, na tradução de bases de metáforas em inglês para o português. Propomos a abordagem Meta4BR, composta por dois componentes: um vertical, que avalia a fidelidade da tradução por meio de retrotradução, e um horizontal, que compara as saídas de diferentes modelos para analisar o grau de concordância entre eles. O objetivo é verificar se a retrotradução oferece uma medida indireta de qualidade e se a convergência entre modelos pode indicar a adequação das traduções. A Figura 1 apresenta a arquitetura da abordagem, detalhada nas seções subsequentes.

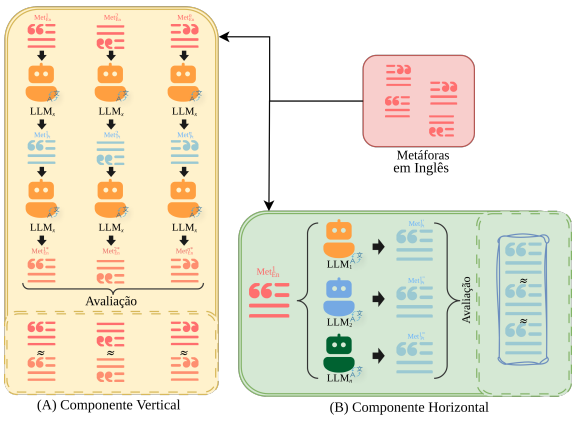

Figura 1. Ilustração do Meta4BR e seus componentes vertical e horizontal.

3.1. Componente vertical: Retrotradução

O componente vertical de avaliação visa atestar a qualidade direta da tradução automática de metáforas com LLMs. Considerando o elevado volume de textos comumente asso-

ciado aos sistemas de tradução automática, é prática recorrente a utilização de métricas automáticas. Contudo, muitas dessas métricas pressupõem a existência de um padrão-ouro de referência [Lin 2004, Zhang et al. 2020, Sellam et al. 2020, Rei et al. 2022], o que representa um entrave em contextos carentes de dados anotados — como é o caso deste trabalho. Dessa forma, a estratégia adotada na avaliação vertical consiste em traduzir os textos da língua de origem (inglês) para a língua-alvo (português) e, em seguida, realizar a retrotradução (*backtranslation*) do conteúdo gerado de volta ao idioma original, utilizando cada modelo avaliado. Assume-se, assim, o texto original como padrão-ouro, sendo a versão retrotraduzida o alvo de comparação. Com esse procedimento, buscamos mensurar a capacidade dos modelos em preservar, entre os idiomas, as informações essenciais contidas nas peças textuais. Abaixo, formalizamos a abordagem.

Formalização Considere um texto T_{orig} na língua original. Um modelo M realiza a tradução para a língua alvo, produzindo $T_{\text{alvo}} = M(T_{\text{orig}})$. Em seguida, aplica-se o modelo novamente para realizar a retrotradução, obtendo-se $T'_{\text{orig}} = M(T_{\text{alvo}})$. A avaliação da preservação de informação entre as línguas pode então ser realizada por meio de uma função de similaridade $S(\cdot, \cdot)$ entre o texto original e sua retrotradução por Qualidade$(M) = S(T_{\text{orig}}, T'_{\text{orig}})$, em que S pode representar métricas como BLEU [Papineni et al. 2002], BERTScore [Zhang et al. 2020] ou COMET [Rei et al. 2022]. Neste trabalho, especificamente, consideramos o inglês como língua original e o português como língua alvo.

3.2. Componente horizontal: Convergência entre Modelos

Para tornar a avaliação das traduções mais robusta na ausência de um padrão-ouro curado, incorporamos o componente horizontal de avaliação. O propósito dessa abordagem é examinar a convergência entre as traduções geradas por diferentes LLMs. Aqui, conjecturamos que, se traduções produzidas por modelos robustos e distintos convergirem para textos semelhantes, isso constitui um indício de plausibilidade e estabilidade no processo tradutório. Para quantificar essa convergência, empregamos medidas estatísticas de similaridade entre observações independentes. Em particular, utilizamos o Coeficiente de Correlação Intraclasse (ICC) [Shrout and Fleiss 1979], tradicionalmente adotado para mensurar a consistência e a confiabilidade entre avaliadores humanos, e aqui adaptado para capturar o grau de concordância entre as saídas geradas pelos diferentes modelos. Especificamente, assumimos os próprios resultados das métricas automáticas como indicadores para cada modelo. A formalização do processo é descrita a seguir.

Formalização Seja $\mathcal{T} = \{T_1^{\text{orig}}, \ldots, T_n^{\text{orig}}\}$ o conjunto de textos na língua original, e $\mathcal{M} = \{M_1, \ldots, M_k\}$ o conjunto de modelos avaliados. Cada modelo $M_j \in \mathcal{M}$ gera uma tradução T_{ij}^{alvo} na língua alvo para o texto T_i^{orig}, com $i \in [1, n]$ e $j \in [1, k]$. Aplicamos uma métrica automática $S(\cdot, \cdot)$ para as traduções T_{ij}^{alvo} geradas, da mesma forma que o componente vertical, considerando as retrotraduções. Para cada item T_i^{orig}, obtemos um vetor de avaliações, com métricas aplicadas a cada modelo, $\boldsymbol{y}_i = \left(S\left(T_i^{\text{orig}},\ T_{ij}^{\text{alvo}} \right) \right)_{j=1}^{k}$.

Com base nessas avaliações, calculamos o ICC, que mede a consistência entre as traduções produzidas pelos modelos para os mesmos textos.

4. Resultados

4.1. Configurações experimentais

Modelos Selecionamos nove modelos, dois LLMs comerciais com reconhecido desempenho geral, GPT-4o-mini [OpenAI 2024] e Gemini 2.0 [Google-DeepMind 2024] (versão Flash-Lite, para equilibrar recursos e desempenho), quatro LLMs mais acessíveis, com desempenho de destaque para seu porte e capacidades multilíngues: Gemma-3-12B [Gemma-Team 2025], Llama 3.1-8B [MetaAI 2024], Ministral-8B [MistralAI 2024] e Qwen 2.5-7B [Alibaba 2025], e três modelos específicos para tradução: GemmaX2-28-9B [Cui et al. 2025], Opus-MT-en-ROMANCE [Tiedemann and Thottingal 2020] e NLLB-200-3.3B [NLLB-Team et al. 2022]. Dessa forma, podemos analisar como diversos tipos de modelos se comportam. GPT-4o e Gemini-2.0 foram aplicados em suas APIs públicas oficiais e os demais foram instanciados com o *framework Transformers* [Wolf et al. 2020] em uma GPU NVIDIA RTX4090. Configurações incluem `temperature` = 1,0, `top_p` = 0,95 e `max_tokens` = 200.

Prompt Para os LLMs não especializados em tradução, aplicamos *prompts* simples e diretos, alinhados com as instruções praticadas na literatura [Cui et al. 2025]. No entanto, dado o contexto metafórico, optamos por incluir, em uma versão do *prompt*, a informação sobre a possível presença de metáforas. O objetivo foi "cientificar" os modelos sobre a necessidade de um cuidado extra na tradução e, assim, avaliar o impacto dessa instrução na produção final. Já os modelos especializados em tradução foram utilizados em suas configurações de instrução padrão. A Figura 2 exibe os *prompts*.

Prompt 1

Traduzir a frase {`frase`} do inglês para o português. Apenas escreva a frase traduzida, nada além disso.

Prompt 2

Traduzir a frase {`frase`} do inglês para o português. Apenas escreva a frase traduzida, nada além disso. A frase pode ou não conter metáfora

Figura 2. *Prompts* utilizados nas traduções.

Bases de dados Foram selecionadas duas bases de dados para os experimentos. A primeira, denominada `manual_data`[2], contém 501 exemplos anotados como metáforas conceituais e 217 exemplos classificados como não-metafóricos. Os exemplos dessa base são curtos, dada a sua natureza, variando entre três e cinco palavras. A segunda, denominada `newsnet` [Joseph et al. 2023], é composta por 2.592 exemplos, dos quais 1.410 são anotados como metáforas. Este conjunto consiste em manchetes de notícias rotuladas manualmente quanto à presença ou ausência de verbos metafóricos, estando, portanto, mais alinhado ao conceito de metáfora linguística.

Métricas Para a análise do **Componente Vertical**, foram empregadas métricas amplamente utilizadas na literatura [Castilho and Caseli 2024], como BLEU [Papineni et al. 2002] e ROUGE [Lin 2004] (baseadas em sobreposição lexical), BERTScore [Zhang et al. 2020] e BLEURT [Sellam et al. 2020] (baseadas

[2]Hub do Hugging Face: Sasidhar1826/manual_data_on_metaphors

em representações semânticas), bem como métricas usadas em tradução, COMET-22 [Rei et al. 2022], KIWI-XL [Rei et al. 2023] e XCOMET-XL [Guerreiro et al. 2024], que consideram aspectos semânticos e contextuais. Para o **Componente Horizontal**, foi utilizado o ICC3k, uma variante do ICC [Shrout and Fleiss 1979] que mede a consistência entre avaliadores fixos avaliando os mesmos itens. No contexto deste trabalho, ele quantifica o grau de concordância das medidas obtidas por diferentes LLMs.

4.2. Resultados Quantitativos

As Tabelas 1,2,3 e 4 apresentam os resultados para os conjuntos `manual_data` e `newsmet`, com modelos genéricos, modelos ajustados para tradução e o componente de retrotradução. Nas tabelas, ✓ indica sentenças com metáfora e ✗, sem metáfora; os desvios-padrão acompanham as médias. **Negrito** indica os melhores valores por modelo.

Tabela 1. Métricas da base `manual_data` com *prompts* 1 (p1) e 2 (p2).

Modelo	Met?	BLEU		ROUGE		BERTScore		BLEURT		COMET22		Kiwi-XL		XComet-XL	
		p1	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*
GPT-4o-mini	✓	**0,258** (0,404)	0,222 (0,377)	**0,665** (0,364)	0,653 (0,375)	**0,889** (0,112)	0,882 (0,112)	**0,380** (0,793)	0,348 (0,826)	**0,822** (0,179)	0,810 (0,182)	**0,780** (0,306)	**0,765** (0,318)	**0,948** (0,085)	0,937 (0,103)
	✗	**0,831** (0,308)	**0,838** (0,287)	**0,939** (0,112)	**0,941** (0,110)	**0,990** (0,020)	**0,991** (0,018)	1,000 (0,194)	1,018 (0,112)	**0,961** (0,037)	**0,964** (0,026)	1,010 (0,124)	1,008 (0,130)	0,997 (0,017)	**0,998** (0,009)
Gemini-2.0	✓	0,102 (0,238)	0,106 (0,238)	0,539 (0,397)	0,553 (0,407)	0,844 (0,123)	0,845 (0,125)	0,111 (0,879)	0,138 (0,902)	0,735 (0,199)	0,740 (0,201)	0,715 (0,387)	0,698 (0,393)	0,877 (0,161)	0,873 (0,173)
	✗	0,807 (0,317)	0,775 (0,332)	0,929 (0,120)	0,919 (0,125)	0,988 (0,022)	0,987 (0,023)	**1,001** (0,157)	0,988 (0,172)	**0,961** (0,031)	0,958 (0,035)	**1,021** (0,102)	**1,018** (0,106)	**0,998** (0,006)	0,997 (0,009)
Gemma-3-12B	✓	0,097 (0,225)	0,096 (0,220)	0,545 (0,400)	0,537 (0,392)	0,834 (0,118)	0,829 (0,113)	0,107 (0,900)	0,108 (0,871)	0,734 (0,198)	0,729 (0,193)	0,702 (0,373)	0,699 (0,386)	0,902 (0,141)	0,900 (0,137)
	✗	0,791 (0,312)	0,803 (0,310)	0,918 (0,129)	0,925 (0,123)	0,986 (0,023)	0,988 (0,022)	0,994 (0,152)	0,995 (0,167)	0,959 (0,032)	0,960 (0,032)	1,006 (0,126)	1,011 (0,110)	**0,998** (0,009)	**0,998** (0,010)
Llama-3.1-8B	✓	0,075 (0,194)	0,062 (0,173)	0,504 (0,359)	0,435 (0,366)	0,822 (0,101)	0,798 (0,103)	0,031 (0,839)	-0,138 (0,840)	0,715 (0,184)	0,677 (0,188)	0,677 (0,385)	0,672 (0,394)	0,892 (0,139)	0,881 (0,147)
	✗	0,701 (0,344)	0,672 (0,363)	0,878 (0,154)	0,869 (0,153)	0,979 (0,029)	0,976 (0,029)	0,939 (0,207)	0,917 (0,235)	0,948 (0,042)	0,944 (0,048)	1,007 (0,121)	0,994 (0,143)	0,996 (0,017)	0,994 (0,023)
Ministral-8B	✓	0,090 (0,229)	0,085 (0,210)	0,571 (0,352)	0,561 (0,358)	0,852 (0,106)	0,846 (0,105)	0,170 (0,808)	0,145 (0,809)	0,760 (0,174)	0,750 (0,173)	0,714 (0,357)	0,696 (0,367)	0,925 (0,107)	0,917 (0,116)
	✗	0,745 (0,338)	0,688 (0,380)	0,899 (0,145)	0,875 (0,180)	0,980 (0,030)	0,974 (0,046)	0,935 (0,264)	0,895 (0,334)	0,946 (0,060)	0,944 (0,058)	1,003 (0,114)	0,988 (0,172)	0,995 (0,021)	0,994 (0,020)
Qwen-2.5-7B	✓	0,161 (0,341)	**0,257** (0,404)	0,517 (0,368)	**0,719** (0,334)	0,849 (0,115)	**0,912** (0,103)	0,045 (0,840)	**0,490** (0,751)	0,750 (0,189)	**0,834** (0,166)	0,739 (0,319)	0,753 (0,309)	0,906 (0,127)	**0,947** (0,092)
	✗	0,756 (0,357)	0,804 (0,306)	0,896 (0,179)	0,923 (0,128)	0,978 (0,053)	0,987 (0,023)	0,912 (0,356)	0,995 (0,168)	0,946 (0,060)	0,957 (0,047)	0,956 (0,233)	1,013 (0,117)	0,989 (0,033)	0,998 (0,011)

Tabela 2. Resultados em `manual_data` com modelos ajustados para tradução.

Modelo	BLEU		ROUGE		BERTScore		BLEURT		COMET22		Kiwi-XL		XComet-XL	
	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
GemmaX2-28-9B	**0,305** (0,439)	**0,694** (0,327)	**0,760** (0,305)	**0,867** (0,143)	**0,933** (0,088)	**0,975** (0,028)	**0,647** (0,633)	**0,978** (0,147)	**0,881** (0,137)	**0,951** (0,040)	0,692 (0,388)	**1,010** (0,122)	**0,966** (0,053)	**0,998** (0,010)
Opus-mt-en-ROMANCE	0,231 (0,391)	0,548 (0,354)	0,749 (0,275)	0,810 (0,149)	0,928 (0,082)	0,962 (0,030)	0,614 (0,583)	0,864 (0,307)	0,862 (0,136)	0,931 (0,054)	**0,700** (0,338)	0,977 (0,162)	0,952 (0,088)	0,995 (0,018)
NLLB-200-3.3B	0,070 (0,191)	0,399 (0,360)	0,464 (0,358)	0,720 (0,214)	0,799 (0,112)	0,935 (0,056)	-0,265 (0,887)	0,616 (0,518)	0,667 (0,188)	0,885 (0,111)	0,544 (0,414)	0,903 (0,258)	0,825 (0,205)	0,959 (0,113)

As Tabelas 1 e 3 indicam ausência de variação significativa entre os *prompts*. Comparando as quatro tabelas, os modelos genéricos superam os ajustados, com exceções pontuais para o GemmaX2 e o Opus-mt. Embora os modelos comerciais tenham, em geral, melhor desempenho, a diferença para modelos menores não é expressiva.

Nas métricas léxicas para o `manual_data`, há distinção significativa entre sentenças com e sem metáforas, tendência não observada no `newsmet`. Embora essa diferença possa refletir características intrínsecas das bases, como discutido na Seção 4.3, pode ser necessário revisar as expressões da base conceitual. BLEURT e Kiwi-XL reforçam essa distinção, enquanto BERTScore e métricas baseadas em COMET apresentam valores elevados e pouca diferenciação entre classes. No BLEURT, apenas dois valores, nas quatro tabelas, são inferiores a zero, indicando saídas potencialmente inaceitáveis segundo [Sellam et al. 2020]. Em geral, os resultados apontam para o potencial da retrotradução como estratégia para geração de metáforas em português.

Tabela 3. Métricas da base `newsmet` com *prompts* 1 (p1) e 2 (p2).

Modelo	Met?	BLEU		ROUGE		BERTScore		BLEURT		COMET22		Kiwi-XL		XComet-XL	
		p1	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*
GPT-4o-mini	✓	0,211 (0,256)	0,206 (0,247)	0,781 (0,147)	0,773 (0,149)	0,931 (0,034)	0,928 (0,033)	0,731 (0,242)	0,717 (0,244)	0,885 (0,053)	0,883 (0,052)	**0,918** (0,128)	**0,921** (0,118)	**0,978** (0,034)	**0,978** (0,037)
	✗	0,280 (0,273)	0,271 (0,264)	0,788 (0,141)	0,785 (0,138)	0,935 (0,032)	0,932 (0,030)	0,733 (0,251)	0,719 (0,250)	0,895 (0,050)	0,893 (0,049)	**0,923** (0,132)	**0,927** (0,122)	**0,978** (0,043)	**0,979** (0,036)
Gemini-2.0	✓	**0,239** (0,281)	**0,236** (0,279)	**0,793** (0,147)	**0,789** (0,146)	**0,942** (0,033)	**0,940** (0,034)	**0,737** (0,263)	**0,735** (0,261)	**0,890** (0,056)	**0,888** (0,057)	0,904 (0,141)	0,897 (0,149)	0,974 (0,045)	0,971 (0,046)
	✗	**0,306** (0,296)	**0,303** (0,295)	**0,806** (0,140)	**0,797** (0,149)	**0,946** (0,032)	**0,943** (0,034)	**0,752** (0,272)	**0,738** (0,292)	**0,901** (0,052)	**0,899** (0,057)	0,916 (0,143)	0,914 (0,141)	**0,978** (0,042)	0,976 (0,046)
Gemma-3-12B	✓	0,196 (0,251)	0,192 (0,249)	0,757 (0,155)	0,756 (0,152)	0,934 (0,035)	0,933 (0,035)	0,702 (0,261)	0,702 (0,261)	0,880 (0,059)	0,880 (0,059)	0,907 (0,139)	0,912 (0,130)	0,974 (0,043)	0,974 (0,042)
	✗	0,243 (0,268)	0,239 (0,265)	0,761 (0,151)	0,760 (0,151)	0,937 (0,034)	0,936 (0,033)	0,706 (0,267)	0,706 (0,267)	0,889 (0,057)	0,888 (0,058)	0,921 (0,130)	0,920 (0,130)	0,977 (0,042)	0,976 (0,043)
Llama-3.1-8B	✓	0,220 (0,242)	0,186 (0,226)	0,706 (0,169)	0,686 (0,173)	0,916 (0,038)	0,908 (0,038)	0,543 (0,358)	0,512 (0,361)	0,863 (0,074)	0,857 (0,073)	0,822 (0,209)	0,836 (0,195)	0,931 (0,097)	0,933 (0,095)
	✗	0,250 (0,247)	0,210 (0,236)	0,726 (0,161)	0,699 (0,169)	0,920 (0,037)	0,912 (0,038)	0,581 (0,331)	0,535 (0,350)	0,873 (0,068)	0,867 (0,071)	0,856 (0,175)	0,860 (0,181)	0,945 (0,083)	0,943 (0,085)
Ministral-8B	✓	0,209 (0,255)	0,193 (0,242)	0,692 (0,176)	0,677 (0,176)	0,918 (0,045)	0,911 (0,046)	0,536 (0,383)	0,498 (0,397)	0,863 (0,079)	0,859 (0,078)	0,827 (0,230)	0,820 (0,236)	0,936 (0,102)	0,931 (0,104)
	✗	0,242 (0,261)	0,225 (0,252)	0,697 (0,171)	0,694 (0,176)	0,921 (0,042)	0,916 (0,043)	0,537 (0,376)	0,527 (0,374)	0,871 (0,074)	0,869 (0,076)	0,848 (0,219)	0,854 (0,212)	0,945 (0,084)	0,946 (0,090)
Qwen-2.5-7B	✓	0,185 (0,235)	0,170 (0,227)	0,673 (0,167)	0,661 (0,167)	0,918 (0,043)	0,913 (0,042)	0,542 (0,374)	0,524 (0,375)	0,859 (0,078)	0,856 (0,075)	0,831 (0,216)	0,841 (0,204)	0,934 (0,104)	0,939 (0,092)
	✗	0,227 (0,256)	0,201 (0,242)	0,693 (0,165)	0,685 (0,168)	0,925 (0,038)	0,919 (0,041)	0,586 (0,344)	0,571 (0,339)	0,871 (0,073)	0,867 (0,069)	0,868 (0,183)	0,865 (0,189)	0,948 (0,089)	0,951 (0,082)

Tabela 4. Resultados em `newsmet` para modelos ajustados para tradução.

Modelo	BLEU		ROUGE		BERTScore		BLEURT		COMET22		Kiwi-XL		XComet-XL	
	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
GemmaX2-28-9B	0,201 (0,274)	0,301 (0,315)	**0,797** (0,152)	**0,810** (0,140)	**0,952** (0,036)	**0,957** (0,031)	**0,756** (0,253)	**0,770** (0,246)	**0,893** (0,057)	**0,906** (0,047)	**0,915** (0,135)	**0,923** (0,128)	**0,980** (0,049)	**0,983** (0,035)
Opus-mt-en-ROMANCE	**0,219** (0,277)	**0,311** (0,306)	0,784 (0,142)	0,802 (0,136)	0,948 (0,034)	0,953 (0,034)	0,597 (0,400)	0,636 (0,362)	0,851 (0,087)	0,870 (0,083)	0,759 (0,246)	0,791 (0,234)	0,929 (0,096)	0,936 (0,094)
NLLB-200-3.3B	0,087 (0,174)	0,115 (0,191)	0,569 (0,225)	0,582 (0,223)	0,882 (0,070)	0,887 (0,070)	0,231 (0,574)	0,263 (0,572)	0,774 (0,135)	0,793 (0,134)	0,685 (0,327)	0,729 (0,308)	0,831 (0,215)	0,851 (0,208)

As Tabelas 5 e 6 apresentam os resultados de ICC3k, que mede a concordância entre os modelos na avaliação horizontal. Melhores valores de *prompts* estão em **negrito**. Observa-se que a variação de *prompt* não afetou significativamente a concordância.

Para metáforas conceituais, BLEU, ROUGE e Kiwi-XL mostraram alta concordância nas sentenças com metáforas, sugerindo que os LLMs seguem padrões lexicais comuns, provavelmente baseados em traduções ou formulações frequentes. Já as métricas semânticas exibiram menor concordância: BERTScore e BLEURT com níveis

moderados e XCOMET-XL com acentuada divergência, indicando maior variabilidade na preservação do sentido metafórico.

No caso das metáforas linguísticas (Tabela 6), o ICC3k apontou alta consistência entre os modelos tanto nas métricas lexicais quanto nas semânticas gerais, possivelmente favorecida pelo contexto envolvente da metáfora. No entanto, a concordância moderada em XCOMET-XL sugere divergências nos aspectos mais finos de fidelidade de sentido. Esse resultado reforça o caráter mais sensível e contextual dessa métrica, destacando a necessidade de análises mais detalhadas para avaliar a qualidade das traduções metafóricas.

Tabela 5. Intraclass Correlation Coefficient (ICC3k) no *dataset* `manual_data`.

Agrupamento	BLEU		ROUGE		BERTScore		BLEURT		COMET22		Kiwi-XL		XComet-XL	
	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
p1+especializados	**0,841**	**0,932**	**0,811**	**0,889**	**0,682**	**0,903**	**0,508**	**0,879**	**0,723**	**0,903**	0,852	0,876	0,249	0,802
p2+especializados	0,821	0,928	0,790	0,882	0,678	0,899	0,491	**0,879**	0,719	0,901	**0,861**	**0,887**	**0,254**	**0,807**
Todos os modelos	0,856	0,940	0,830	0,907	0,725	0,918	0,540	0,902	0,745	0,920	0,879	0,903	0,259	0,835

Tabela 6. Intraclass Correlation Coefficient (ICC3k) no *dataset* `newsmet`.

Agrupamento	BLEU		ROUGE		BERTScore		BLEURT		COMET22		Kiwi-XL		XComet-XL	
	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
p1+especializados	**0,885**	**0,888**	**0,851**	**0,851**	**0,826**	**0,829**	**0,804**	**0,811**	**0,856**	**0,862**	**0,804**	0,811	**0,663**	**0,683**
p2+especializados	0,881	0,884	0,848	**0,851**	0,824	0,827	0,799	0,809	**0,856**	**0,862**	0,802	**0,814**	0,661	0,682
Todos os modelos	0,905	0,906	0,872	0,873	0,848	0,850	0,829	0,836	0,873	0,878	0,831	0,840	0,691	0,709

4.3. Avaliação Humana

Especificação Com base em trabalhos anteriores de avaliação humana de traduções [Freitag et al. 2021] e de traduções de metáforas em outros idiomas [Wang et al. 2024], adotamos oito critérios[3] para a inspeção manual: (1) **Fluência**: A tradução soa natural e está gramaticalmente correta; (2) **Inteligibilidade**: O sentido metafórico é claro, mesmo que adaptado; (3) **Compreensão**: O leitor entende a metáfora, sua intenção e possíveis nuances culturais ou emocionais; (4) **Equivalência Total**: A tradução mantém o sentido literal e o metafórico original; (5) **Equivalência Parcial**: A metáfora é preservada, mas com imagem ou forma diferente; (6) **Não-equivalência**: A tradução perde o tom metafórico, alterando o sentido literal e contextual; (7) **Tradução incorreta**: A palavra-alvo foi mal traduzida, mudando seu sentido literal e contextual; (8) **Não-tradução**: Tradução totalmente incorreta ou sem relação com o texto original.

Para os quesitos 1 a 3, utilizamos a escala de Likert de 5 pontos, na qual 5 corresponde a "Muito Bom" e 1 a "Muito Ruim". Os demais quesitos são do tipo "Sim" ou "Não". Cada segmento é constituído por um trecho na língua original e sua respectiva tradução. Nos casos em que o segmento não contém uma expressão metafórica, apenas os quesitos referentes às categorias de erro (7 e 8) devem ser preenchidos.

Respostas Consideramos um subconjunto de 70 expressões incluídas na base de metáforas conceituais (`common`[4]). Esse conjunto contém 10 expressões anotadas como não-metafóricas. Entretanto, após uma inspeção pelos autores deste artigo, decidimos

[3] As definições são mais longas e contém exemplos, que aqui foram omitidos por falta de espaço.

[4] Hub no Hugging Face: Sasidhar1826/common_metaphors_detection_dataset

que quatro expressões também poderiam ser consideradas como contendo expressões metafóricas[5]. Como são apenas seis expressões e todas elas foram anotadas com o valor de "Não" para os quesitos 8–10, exibiremos apenas os valores para as demais 64 expressões na Tabela 7. Os valores foram coletados a partir da anotação de uma das autoras do artigo, especialista em linguística, e revisados por outra autora, sem casos de discordância.

Tabela 7. Contabilização da inspeção humana das traduções, incluindo casos incorretos e de não-tradução. Equivalência metafórica: total, parcial e não-equivalência. Níveis: fluência, inteligibilidade e compreensão.

Modelo	Inc.	NT	Equivalência			Nível 1			Nível 2			Nível 3			Nível 4			Nível 5		
			Tot	Par	Não	Flu	Int	Com	Flu	Int	Com	Flu	Int	Com	Flu	Int	Com	Flu	Int	Com
Gemini-2.0	0	0	47	7	10	2	2	2	8	8	8	0	0	0	17	9	8	37	45	46
GemmaX2	2	0	43	9	10	3	3	3	9	9	9	0	0	0	9	10	9	43	42	43
Gemma-3	0	1	45	8	11	1	1	1	9	9	9	1	1	1	8	8	8	45	45	45
GPT-4o	0	0	45	8	11	2	2	2	9	9	9	0	0	0	16	10	9	37	43	44
Llama-3.1	3	0	40	11	9	4	4	4	9	10	10	1	0	0	12	9	9	38	41	41
Ministral	5	1	41	11	12	4	4	4	9	9	9	0	0	0	28	10	10	23	41	41
NLLB-200	7	0	39	11	13	7	7	7	7	7	7	0	0	0	18	9	9	32	41	41
Opus-MT	3	0	42	9	13	3	3	3	10	10	10	0	0	0	28	10	9	23	41	42
Qwen-2.5	7	1	39	11	14	8	8	8	6	6	6	1	0	0	43	12	10	6	38	40

O modelo Gemma-3 apresentou o melhor desempenho no nível 5 da escala de Likert para os critérios 1 a 3, além de liderar nos demais. Os modelos concentraram-se nos níveis 4 e 5, indicando boa capacidade de captar metáforas conceituais. Nos casos de discordância com a anotação original, as traduções foram literais, como "*She writes the office*" traduzida por "Ela escreve no escritório" ou "Ela escreve *The Office*", em vez de sentidos esperados como "projetar" ou "customizar". Um caso notável é "*Time has a mountain*", traduzido literalmente como "O tempo tem uma montanha" por todos os modelos, sem sentido claro em português ou inglês, o que pode apontar falhas na base original. Erros ou omissões de tradução foram raros, com predominância de equivalência total, seguida por equivalência parcial e não equivalência.

5. Conclusões

Este artigo apresentou a Meta4BR, uma abordagem baseada em retrotradução e saídas de diferentes LLMs para avaliar a qualidade da tradução de sentenças com metáforas do inglês para o português. Resultados obtidos em duas bases, utilizando nove modelos, métricas lexicais e semânticas, análise de concordância e avaliação humana indicam o potencial de criar bases de metáforas a partir de recursos existentes em outras línguas, como o inglês. Mas também indicam a necessidade de uma análise refinada em casos de discordância entre modelos e de identificação de traduções em sentido literal. Ademais, a geração de várias traduções por diferentes modelos permite utilizar as métricas para selecionar a opção mais adequada ao contexto.

Trabalhos futuros envolvem novos experimentos com diferentes bases e *prompts*, análise de casos de alta e baixa concordância (tradução literal e perda de sentido) e a comparação entre anotações humanas e automáticas na detecção de metáforas.

Agradecimentos

Os autores agradecem o apoio financeiro do CNPq, dos INCT IAIA e TILD-IAR e da FAPERJ (processos SEI-260003/002930/2024 e SEI-260003/000614/2023).

[5]As expressões são: *She/The chef writes the office/restaurant* e *The child/My friend drinks the market/store*. Conjecturamos que o desvio no sentido pode ter sido devido à falta de preposição.

Referências

Alibaba (2025). Qwen2.5 Technical Report.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Boisson, J., Espinosa-Anke, L., and Camacho-Collados, J. (2023). Construction Artifacts in Metaphor Identification Datasets. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590, Singapore. Association for Computational Linguistics.

Boisson, J., Mehmood, A., and Camacho-Collados, J. (2025). METAPHORSHARE: A dynamic collaborative repository of open metaphor datasets. In Dziri, N., Ren, S. X., and Diao, S., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 509–521, Albuquerque, New Mexico. Association for Computational Linguistics.

Boldarine, A. C. (2024). Uso de metáforas por aprendizes brasileiros bilíngues de inglês.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Castilho, S. and Caseli, H. d. M. (2024). Tradução automática: Abordagens e avaliação. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 23. BPLN, 3 edition.

Cui, M., Gao, P., Liu, W., Luan, J., and Wang, B. (2025). Multilingual Machine Translation with Open Large Language Models at Practical Scale: An Empirical Study. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

Deignan, A. (2016). From linguistic to conceptual metaphors. In *The Routledge handbook of metaphor and language*, pages 120–134. Routledge.

Freitag, M., Foster, G. F., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Trans. Assoc. Comput. Linguistics*, 9:1460–1474.

Gemma-Team (2025). Gemma 3 technical report.

Glucksberg, S. and Keysar, B. (1993). How metaphors work. *Metaphor and thought*, 2:401–424.

Goatly, A. (1997). *The language of metaphors*. Routledge.

Google-DeepMind (2024). Introducing Gemini 2.0: our new AI model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/. Acesso em 10 jun. 2025.

Guerreiro, N. M., Rei, R., Stigt, D. v., Coheur, L., Colombo, P., and Martins, A. F. T. (2024). xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Joseph, R., Liu, T., Ng, A. B., See, S., and Rai, S. (2023). NewsMet : A 'do it all' dataset of contemporary metaphors in news headlines. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10090–10104, Toronto, Canada. Association for Computational Linguistics.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kabra, A., Liu, E., Khanuja, S., Aji, A. F., Winata, G., Cahyawijaya, S., Aremu, A., Ogayo, P., and Neubig, G. (2023). Multi-lingual and Multi-cultural Figurative Language Understanding. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Kövecses, Z. (2005). *Metaphor in culture: Universality and variation*. Cambridge university press.

Kövecses, Z. (2010). Metaphor, language, and culture. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 26:739–757.

Krennmayr, T. and Steen, G. (2017). *VU Amsterdam Metaphor Corpus*, pages 1053–1071. Springer Netherlands, Dordrecht.

Labarta Postigo, M. (2023). 'Life is baseball' vs. 'A vida é futebol': la traducción de metáforas deportivas del inglés al portugués de Portugal y al portugués de Brasil. *Sendebar*, 34:147–161.

Lakoff, G. and Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, T., Wang, Y., Wang, Z., and Yu, H. (2025). Exploring cognitive effort and divergent thinking in metaphor translation using eye-tracking and EEG technology. *Scientific Reports*, 15(1):18177.

Ma, W. (2025). Optimizing English translation processing of conceptual metaphors in big data technology texts. *Advances in Continuous and Discrete Models*, 2025(1):36.

Massey, G. (2021). Re-framing conceptual metaphor translation research in the age of neural machine translation: Investigating translators' added value with products and processes. *Training, Language and Culture*, 5(1):37–56.

MetaAI (2024). The Llama 3 Herd of Models.

MistralAI (2024). Ministral-8B. Acesso em 09 jun. 2025.

Morrison, J., Na, C., Fernandez, J., Dettmers, T., Strubell, E., and Dodge, J. (2025). Holistically evaluating the environmental impact of creating language models. In *The 13th International Conference on Learning Representations*.

NLLB-Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation.

OpenAI (2024). Gpt-4o System Card.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Postigo, M. L. (2024). ChatGPT and MT-Systems: advantages and limitations when translating english to spanish and portuguese. *Lengua y Habla*, 28(0):369–390.

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., and Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Rei, R., De Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., and Martins, A. F. (2022). Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Rei, R., Guerreiro, N. M., Pombal, J., van Stigt, D., Treviso, M., Coheur, L., C. de Souza, J. G., and Martins, A. (2023). Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Sanchez-Bayona, E. and Agerri, R. (2024). Meta4XNLI: A Crosslingual Parallel Corpus for Metaphor Detection and Interpretation.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Shao, Y., Yao, X., Qu, X., Lin, C., Wang, S., Huang, W., Zhang, G., and Fu, J. (2024). CMDAG: A Chinese metaphor dataset with annotated grounds as CoT for boosting metaphor generation. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3357–3366, Torino, Italia. ELRA and ICCL.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Shutova, E. (2017). Annotation of linguistic and conceptual metaphor. *Handbook of linguistic annotation*, pages 1073–1100.

Steen, G. (2010). A Method for Linguistic Metaphor Identification: From MIP to MIPVU.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In Martins, A., Moniz, H., Fumega, S., Martins, B., Batista, F., Coheur, L., Parra, C., Trancoso, I., Turchi, M., Bisazza, A., Moorkens, J., Guerberof, A., Nurminen, M., Marg, L., and Forcada, M. L., editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Tong, X., Choenni, R., Lewis, M., and Shutova, E. (2024). Metaphor Understanding Challenge Dataset for LLMs. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, S., Zhang, G., Wu, H., Loakman, T., Huang, W., and Lin, C. (2024). MMTE: Corpus and Metrics for Evaluating Machine Translation Quality of Metaphorical Language. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11343–11358, Miami, Florida, USA. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xia, M., Kong, X., Anastasopoulos, A., and Neubig, G. (2019). Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

Zayed, O., McCrae, J. P., and Buitelaar, P. (2020). Figure Me Out: A Gold Standard Dataset for Metaphor Interpretation. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5810–5819, Marseille, France. European Language Resources Association.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating Text Generation with Bert. In *International Conference on Learning Representations*.