超越句子：基于大语言模型的情境感知机器翻译调查

Ramakrishna Appicharla[1*], **Baban Gain**[1*], **Santanu Pal**[2], **Asif Ekbal**[3]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2]Wipro AI, Lab45, London, UK
[3]School of AI and Data Science, Indian Institute of Technology Jodhpur, India
{ ramakrishnaappicharla,gainbaban, santanu.pal.ju, asif.ekbal } @gmail.com

Abstract

尽管大型语言模型（LLMs）非常流行，但其在机器翻译中的应用相对较少探索，尤其是在上下文感知设置中。本文对使用 LLMs 进行上下文感知翻译的文献进行了综述。现有的工作利用了提示和微调的方法，极少数工作集中于自动后编辑和为上下文感知机器翻译创建翻译代理。我们观察到商业 LLMs（如 ChatGPT 和 Tower LLM）比开源 LLMs（如 Llama 和 Bloom LLMs）取得了更好的结果，并且基于提示的方法可以作为评估翻译质量的良好基准。最后，我们提出了一些值得探索的有趣未来方向 [1] 。

1 介绍

机器翻译（MT）能在不丧失意义的情况下将自然语言句子从一种语言翻译成另一种语言。通过基于神经网络的方法 (Sutskever et al., 2014; Bahdanau et al., 2014)，特别是基于 transformer 的 (Vaswani et al., 2017) 模型，这一技术取得了巨大进步。然而，大多数方法在句子层面进行翻译，并且经常无法处理话语现象（如代词和省略）(Voita et al., 2018; Wang et al., 2023a)，从而导致不一致的翻译。文档级（或上下文感知）翻译 (Maruf and Haffari, 2018; Zhang et al., 2018; Bawden et al., 2018; Agrawal et al., 2018; Voita et al., 2019; Huo et al., 2020; Li et al., 2020; Donato et al., 2021; Maruf et al., 2021) 通过训练模型翻译文档或段落而不是句子来解决这个问题。构建上下文感知神经机器翻译（NMT）模型主要有两种方法：基于串联的 (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Junczys-Dowmunt, 2019; Zhang et al., 2020) 方法和多编码器 (Zhang et al., 2018; Voita et al., 2018; Kim et al., 2019; Ma et al., 2020) 方法。Sun et al. (2022) 已显示标准的 transformer 模型拥有足够的能力有效地处理话语级现象。然而，构建上下文感知 MT 系统的主要限制之一是缺乏文档级 (Zhang et al., 2022a) 平行语料库。

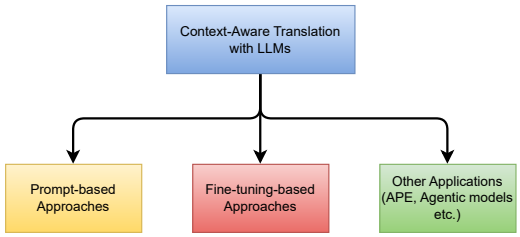

Figure 1: 本研究中所呈现的调查概述。

大型语言模型（LLMs）已经成为构建各种自然语言处理（NLP）应用的新范式，包括机器翻译（MT）。Pang et al. (2025) 使用 LLMs 重新评估了 NMT 模型的六个挑战，并报告说数据稀缺不再是一个显著的挑战，因为 LLMs 是在更大规模的数据上进行预训练的，特别是在文档级翻译方面。他们对德英语对进行了 Llama-2 的微调，并报告说使用句子级数据微调的 LLMs 在较长句子的翻译性能上表现良好。同样，已经有几项工作尝试用 LLMs 改善文档级翻译。这项工作报告了当前有关使用 LLMs 进行上下文感知翻译的研究。已经有一些调查专注于非 LLM 的文档级翻译 (Maruf et al., 2021; Peng et al., 2024) 或 LLM 的句子级翻译 (Lyu et al., 2024a; Gain et al., 2025)。该调查的主要目的是提供 LLMs 的文档级 MT 的详细回顾，这可以帮助构建更高效的模型来处理文档级翻译任务。

论文的剩余部分组织如下：第 2 节和第 3 节分别描述了基于提示和基于微调的方法（参见图 1）。这两部分包含了探索使用 LLMs 进行上下文感知翻译的工作。在第 4 节，我们描述了与上下文感知翻译相关的工作，如 APE、探索上下文使用，并为文档级翻译开发一个代理框架。最后，第 **??** 节对工作进行总结，并提出了一些潜在的未来研究方向。

2 基于提示的方法

提示是一个自然语言的陈述或指令，用于引导大型语言模型（LLM）的输出朝向特定方

[*]px Joint First Authors
[1]简而言之，请参阅表格 1、2 和 3，了解本调查中涉及的工作的概览。

向。提示技术是一种推理时的技巧，并不需要对 LLM 进行任何训练或微调。常用的基于提示的方法来构建上下文感知的机器翻译系统有：（i）零样本提示（ZSP），以及（ii）少样本提示或上下文学习（ICL）。在 ZSP 中，LLM 仅被给予提示（例如，将此句子从英语翻译成德语：<sentence>）以执行特定任务，并且没有提供任何额外信息来说明如何执行该任务。如果底层的 LLM 没有专门为特定任务进行训练，ZSP 可能导致次优的性能。在 ICL 中，LLM 被提供了一些例子（输入-输出对）和提示。这些例子提供了必要的上下文以有效完成任务并提高性能。基于 ZSP 和 ICL 的方法被用来开发使用 LLM 的上下文感知的机器翻译系统。本节主要描述采用不同基于提示的方法（参见表 1 ）来构建和评估上下文感知的机器翻译系统的工作。

Wang et al. (2023b) 通过 ZSP 开展实验以评估 LLM 的语篇建模能力。具体来说，他们设计了三个提示，其中两个提示用于在单个对话回合内翻译文档 [2] ，一个提示用于在多个对话回合中逐句翻译文档。所有实验都是在 ChatGPT [3] 上进行的。他们在 ChatGPT 的在线聊天版本上进行了实验。他们报告说，所有三个提示的表现都同样出色，甚至逐句翻译的提示结果也很有前途，因为 ChatGPT 通常能够在单次聊天会话中很好地记住上下文。他们还报告说，ChatGPT（GPT-4）在语篇级翻译方面优于商业 MT 系统，如 Google 翻译、DeepL Translator 和 Tencent TranSmart。

Hendy et al. (2023) 在 ZSP 和 ICL 设置中对 GPT-3.5（text-davinci-003）进行了实验，以研究上下文如何影响翻译性能。具体来说，在 ZSP 设置中，每篇文档逐句翻译，且上下文长度的数量有所不同（范围为 1-32，其中 1 表示每次只翻译一个句子而没有任何上下文，而 32 表示所有 32 个句子同时以逐句方式翻译）。他们观察到，增加上下文窗口的长度会带来更好的性能。他们还进行了 ICL 实验，使用了五个例子和一个包含 10 个句子的上下文窗口。他们报告称，少样本学习对文档级翻译是多余的，因为源文档提供了足够的上下文以进行有效翻译。

Moslem et al. (2023) 进行了文档级翻译的实验，将模糊匹配（也称为翻译记忆）作为 ICL 的例子。实验在 GPT-3.5（text-davinci-003）以及 Bloom (Workshop et al., 2022) 和 Bloomz (Muennighoff et al., 2023) 大型语言模型上进行，文档被逐句翻译。他们报告称，ICL（最多 5 个例子）的表现优于 ZSP 和随机选择的例子。

Zhang et al. (2023a) 研究了能否将从句子级语料库中选择的示例转移来翻译文档级数据。他们在 GLM-130B (Zeng et al., 2022) LLM 上进行了一次性提示设定的实验，使用从句子级语料库中提取的示例翻译一个包含约四个句子的文档块。结果表明，1 次提示的表现在文档-BLEU 和特定文档的 (Sun et al., 2022) 指标上均优于 ZSP。

Wu and Hu (2023) 对 GPT-3.5 进行了实验，使用 ZSP 对文档逐句翻译。他们报告说，将句子作为多轮对话进行翻译比单独翻译句子效果更好。他们还提示模型在一次对话中进行翻译并修订翻译。然而，这种方法的表现最低，因为 LLM 无法修订结果，因为没有明确或具体的指令来指导修正过程。

Bawden and Yvon (2023) 在 1-shot 设置中对 Bloom (Workshop et al., 2022) 和 Bloomz (Muennighoff et al., 2023) LLM 进行了实验。他们选择了两种类型的例子：随机和之前的对话话语。他们报告说，之前的对话例子给出的分数最好；随机例子选择和之前对话选择的结果差异非常小。目前尚不清楚之前句子的上下文是否比随机上下文更好。这些结果与之前在非 LLM DocMT (Li et al., 2020; Sun et al., 2022; Appicharla et al., 2023, 2024) 模型上的发现一致。

文档级平行语料库的可用性比句子级平行语料库要少，而且在文学翻译中则更少。Karpinska and Iyyer (2023) 为 18 种语言对创建了段落级平行语料库，这些语料库来自公开可获得的小说和短篇小说（某些小说和短篇小说的原文和翻译版本均可用。收集的数据均非合成）。他们在三种不同的设置下对 GPT-3.5（text-davinci-003）进行了包含五个例子的 ICL 测试：逐句翻译、逐段翻译以及逐段翻译但带有句子标记以区分句子之间。他们还进行了 MQM (Freitag et al., 2021) 并用七种不同类型的错误类别 [4] 对翻译进行标注，同时进行人工评估。结果显示，人们更偏爱逐段翻译的设置而非其他输出。然而，常见的错误包括误译和语法错误。

Sia and Duh (2023) 研究了为 ICL 选择的示例如何影响翻译的连贯性。为了研究上下文（在示例中提供的）如何影响文档级翻译，他们比较了五个随机选择的示例与五个之前句子的示例（移动窗口方法）。源句子及其标准翻译作为示例提供，并对文档逐句翻译。所有实验都在

[2]这两个提示是相似的，但其中一个提示包括用于指示句子边界的标签。

[3]实验是在 GPT-3.5 和 GPT-4 版本上进行的。

[4]1. 错译, 2. 语法, 3. 未翻译的词, 4. 不一致, 5. 语域（正式和非正式的语言使用）, 6. 格式（正确使用标点符号）, 7. 增补和省略

Work	LLMs	Language Pairs
Wang et al. (2023b)	GPT-3.5, GPT-4	Zh → En, En → De, En → Ru
Hendy et al. (2023)	ChatGPT, GPT-3.5 (text-davinci-002, 003)	En ↔ De
Moslem et al. (2023)	ChatGPT, GPT-3.5 (text-davinci-003), Bloom, Bloomz	En → { Ar, Zh, Fr, Rw, Es }
Zhang et al. (2023a)	GLM	Zh → En
Wu and Hu (2023)	GPT-3.5	En ↔ Zh
Bawden and Yvon (2023)	Bloom, Bloomz	En ↔ Fr
Karpinska and Iyyer (2023)	GPT-3.5 (text-davinci-003)	{ En, De, Fr, Ru, Cs, Ja } → Pl; { En, De, Fr, Ru, Pl, Zh } → Ja; { De, Fr, Ru, Pl, Ja, Zh } → En
Sia and Duh (2023)	GPT-Neo, Bloom, XGLM	En → { Fr, De, Pt }
Cui et al. (2024a)	Qwen, Baichuan	En ↔ De, En ↔ Zh
Enis and Hopkins (2024)	Claude 3 Opus	Yo → En
Hu et al. (2025)	Qwen2.5, Llama-3.1, GPT-4o-mini	En → { Zh, Cs, De, Hi, Is, Ja, Ru, Es, Uk } ; Cs → Uk; Ja → Zh

Table 1: 本文涵盖的基于提示的方法列表。

GPT-Neo (Black et al., 2021) 、Bloom (Workshop et al., 2022) 、XGLM (Lin et al., 2021) 上进行，他们报告说，与随机选择相比，前五句上下文可以实现最佳结果。他们还观察到，从同一文档中随机选择示例比从文档外选择示例表现更好。然而，这些示例的顺序（例如，保持与文档中相同的顺序与打乱这些句子的顺序）并不影响性能。最后，他们还注意到保持静态上下文（例如选择前五句并在整个文档翻译过程中使用它们）表现不如随机抽样。这表明选择正确的上下文对于高效翻译至关重要，但这些示例的顺序可以有所不同。

Cui et al. (2024a) 提出了一种上下文感知的提示方法，通过 ICL 方法解决由 LLMs 生成的不连贯翻译问题。具体而言，他们提出了一种上下文感知提示方法（CAP）来选择 ICL 的示例。所提出的 CAP 方法包括三个步骤。在第一步，根据句子级注意力分数（动态上下文窗口）提取与当前句子相似的句子。在第二步，总结第一步中提取的上下文，并用于从外部数据库中检索相似的句子。最后，检索到的句子被用作 ICL 的示例。实验在 Baichuan (Yang et al., 2023) 和 Qwen (Bai et al., 2023) LLMs 上进行。他们还将所提出的方法与不同的提示策略进行了比较，比如 ZSP 和 ICL，以及随机示例和前三个句子的组合。所提出的方法取得了最佳结果。然而，ZSP 的性能与所提出的方法也非常相似。在随机上下文的情况下，较大的 LLMs（参数范围在 14-72B）比较小的 LLMs 对随机上下文更稳健，其中示例是从数据库中随机选择的。当模型在零代词翻译（ZPT）任务中进行评估时，ZSP 与其他 ICL 方法相比表现较差。这表明上下文对有效的上下文感知翻译至关重要。

Enis and Hopkins (2024) 在资源匮乏的约鲁巴语-英语语对上进行了 Claude LLM 的实验。具体来说，他们遵循 ICL 将约鲁巴语新闻文档（从 BBC 抓取）逐句翻译成英语。输入给 LLM 的示例由拆分成句子及其对应翻译的文档组成，然后提示 LLM 根据给定的示例逐句翻译另一篇文档。他们报告称，Claude LLM 在自动评分指标（即 BLEU 和 ChrF++）方面表现显著优于传统机器翻译系统，如谷歌翻译和 NLLB (Costa-Jussà et al., 2022) 模型。然而，他们没有对获得的翻译进行深入的上下文感知分析。

Hu et al. (2025) 进行了在多轮对话环境下翻译文档的实验。具体而言，他们将多轮翻译性能与单轮（即整个文档在一个轮次中翻译）和段级翻译（即将文档分成多个段落并独立翻译）进行了比较。相比其他方法，他们观察到单轮翻译表现不佳，因为在翻译长文档时容易出现遗漏错误。多轮翻译设置则由于可以缓存并使用之前轮次的翻译作为上下文而取得了最佳结

果。他们在翻译过程中还将整个源文档提前输入，以进一步提升多轮翻译设置的翻译质量。所有实验均在 GPT-4o-mini (Hurst et al., 2024) 、Llama-3 (Grattafiori et al., 2024) 和 Qwen-2.5 (Yang et al., 2024a) LLMs 上进行。最后，他们也观察到性能可以通过使用 ICL（在实验中使用了三个示例）得到提升。

3 基于微调的方法

通过微调（FT），LLM 准确翻译上下文感知数据的能力可以在基于提示的方法基础上进一步提升。可以基于特定语言对或领域数据对 LLM 进行进一步微调。通常，微调 LLM 以用于特定的下游任务显示了性能的改进 (Brown et al., 2020) ，但需要更多的计算资源。有多项研究集中在通过微调 LLM 来提高上下文感知翻译性能（参见表 2 ）。

Zhang et al. (2023b) 是较早探索使用提示和微调方法在句子和文档级别设置中不同大型语言模型（LLM）翻译性能的工作之一。他们在 GPT-Neo (Black et al., 2021) 、OPT (Zhang et al., 2022b) 、Llama-2 (Touvron et al., 2023b) 、XGLM (Lin et al., 2021) 和 Bloomz (Muennighoff et al., 2023) LLMs 上进行了实验。他们研究了零样本学习（ZSP）和上下文内学习（ICL）设置，并使用 QLoRA (Dettmers et al., 2023) 对 LLM 进行微调。他们在法语到英语的语言对上进行了实验。研究发现，在句子和文档级别的翻译设置中，微调一致地比基于提示的方法取得了更好的结果。此外，他们报告说，使用文档长度为 10 个句子进行微调的模型比使用每个文档 5 或 15 个句子进行训练的其他模型取得了更好的结果。他们注意到，微调多语言大型语言模型（XGLM 和 Bloomz）比使用提示能显著取得更好的结果。

Wu et al. (2024a) 比较了各种非 LLM 模型 (Costa-Jussà et al., 2022; Sun et al., 2022; Wu et al., 2024b) 、闭源模型（如谷歌翻译、GPT-3.5-Turbo 和 GPT-4-Turbo）与开源 LLM，如 Llama-2 (Touvron et al., 2023b) 、Bloom (Workshop et al., 2022) 和经过文档级平行语料库微调的 Vicuna (Chiang et al., 2023) 。他们在一个两步过程中微调 LLMs，首先在单语文档 (Xu et al., 2023) 上训练 LLM，接着是平行文档。他们使用前面的三个句子作为背景。他们采用全文微调（FFT）和低秩适应（LoRA）(Hu et al., 2022) 微调 LLMs。他们在九种语言对上训练了模型。他们报告说在双语场景（每个语言对单独训练时），LoRA 表现优于 FFT，而在多语言场景中，FFT 表现更好。他们还观察到 FFT 需要 1 % 的数据，而 LORA 需要 10 % 才能实现相同的性能。然而，GPT-4-Turbo 在所有语言对中与 NLLB 一起达到了最佳结果，微调的 Bloom-7B 模型在英译其他语言和其他语言译英文对中分别取得了次优的结果。他们还尝试了两种不同的推论方法：(i) 使用先前翻译的句子作为上下文，和 (ii) 独立翻译上下文中的每个句子。他们报告说，使用先前翻译的上下文会导致翻译质量下降，因错误传播所致，而独立翻译每个句子可以获得良好的结果，但计算成本更高。

在句子级别语料库上微调的 LLMs 无法翻译较长的文档（包含超过 512 个标记）。为了有效翻译较长的文档，Li et al. (2024) 提出结合不同长度的句子和文档级别指令来微调 LLMs。文档被分成长度为 $[512, 1024, 1536, 2048]$ 标记的子文档，并与句子级别指令结合起来，创建出翻译混合指令。他们遵循 Alpaca 指令格式 (Taori et al., 2023) 来形成这些翻译指令并微调 Llama-2 (Touvron et al., 2023b) 。在进行从不同语言到英语的翻译时，使用翻译混合指令（子文档长度 1024 设置）微调的模型相比其他微调设置取得了最佳结果。然而，当从英语翻译到其他语言时，基于非 LLM 的 DocNMT 模型 (Li et al., 2023) 在 LLM 模型中取得了最佳结果。所提出的方法还通过更长长度的指令（特别是 1024 和 1536 设置）提高了话语级别翻译精度，在话语现象评估中取得了最佳结果。

在进行大语言模型的微调以实现上下文感知的翻译时，输入由连接的上下文和源句组成。然而，上下文和源句在连接时被赋予相同的优先级。这可能对有效翻译不利，因为源应该比上下文具有更高的优先级。为了解决这个问题，Lyu et al. (2024b) 提出了一个技术（称为解码增强多提示调优），能够让大语言模型区分上下文和源句，从而提升翻译性能。具体来说，模型训练有三个步骤。在第一步中，所有的上下文句子被连接形成一个单一序列以进行编码。在第二步中，当前源句基于从前一步获得的激活进行编码。第一步和第二步得到的激活用于第三步中预测目标序列概率。通过结合上述三个步骤的激活，进一步增强了解码。在每个阶段中都引入了可训练的提示，并且遵循深度提示调优 (Li and Liang, 2021; Liu et al., 2021) 来训练模型。他们在 Llama-2 (Touvron et al., 2023b) 和 Bloomz (Muennighoff et al., 2023) 大语言模型上进行了实验，并将上下文窗口设置为最大 256 个标记。提出的方法显著优于非大语言模型的模型 (Bao et al., 2021; Sun et al., 2022) ，以及训练时使用连接上下文（即没有多阶段训练）和没有上下文（即在没有上下文的情况下进行微调）的大语言模型。他们报告称，提出的方法对上下文变化更具鲁棒性，并且在较长

Work	LLMs	Language Pairs
Zhang et al. (2023b)	GPT-Neo, OPT, Llama-2, Bloomz	Fr → En
Wu et al. (2024a)	GPT-3.5-Turbo, GPT-4-Turbo, Llama-2, Vicuna, Bloom	En ↔ { Ar, De, Fr, It, Ja, Ko, Nl, Ro, Zh }
Li et al. (2024)	Llama-2	En ↔ { Fr, De, Es, Ru, Zh }
Lyu et al. (2024b)	Llama-2, Bloomz	{ Fr, De, Es, Ru, Zh } → En
Wang et al. (2024a)	GPT-3.5, GPT-4, Llama-2	Zh → En; En → { De, Ru }
Kudo et al. (2024)	Llama-2, Mistral	En → Ja, Ja → Zh
Sung et al. (2024)	GPT-4o, Gemma	En ↔ Ko
Yang et al. (2024b)	Llama-3	En ↔ { De, Nl, Pt-Br }
Pombal et al. (2024)	Tower-Chat	En ↔ { Fr, De, Nl, Ko, Pt-Br }
Wu et al. (2024d)	Llama-2	En → Zh
Elshin et al. (2024)	YandexGPT	En → Ru
Zafar et al. (2024)	Llama-3	{ Fr, De } → En
Luo et al. (2024)	Chinese-Llama-2	Zh → En

Table 2: 本文中涵盖的微调方法列表。

上下文输入上表现良好。

Wang et al. (2024a) 构建了基于指令的数据集，用于微调大型语言模型，使其能够进行上下文感知的翻译。具体而言，他们遵循 Alpaca (Taori et al., 2023) 格式来创建指令，并将这些指令添加到现有的文档级语料库中，将其转换为基于指令的语料库。他们使用 GPT-4 生成指令（例如，将以下句子从英文翻译成德文：源句子），并为多种语言（中英文、俄英文和德英文）以及不同领域（新闻、字幕、TED 演讲、议会记录、小说）创建了数据集。在推理过程中，句子被连接到大型语言模型支持的最大长度。他们比较了不同商业机器翻译系统（谷歌翻译、DeepL 和腾讯翻译）、闭源大型语言模型（GPT-3.5、GPT-4）以及开源大型语言模型（Llama-2）的结果。结果显示 GPT-4（具有 8K 输入长度）在其他系统中实现了最佳结果（除了在小说领域）。他们还在准备好的语料库上微调了 Llama (Touvron et al., 2023a)，以研究文档长度的影响。他们发现，使用最大长度为 2048 个标记训练的模型优于使用最大长度为 4096 个标记训练的模型以及句子级模型。他们报告说，在微调过程中变化文档长度能在固定长度文档的模型上获得更好的结果。最后，他们进行了多任务学习的实验，其中辅助任务是文本补全（即根据给定上下文生成后续句子）。他们报告说，多任务学习提高了翻译性能，尤其是在文学等领域。

Kudo et al. (2024) 进行了关于最小贝叶斯风险（MBR）解码 (Eikema and Aziz, 2020) 和基于 LLM 的重新排序以提高翻译性能的实验。他们最初训练了非 LLM 和基于 LLM 的 MT 模型用于句子级翻译。在解码过程中，使用 COMET-22 (Rei et al., 2022) 作为效用函数，通过 MBR 解码选择最佳假设。在 MBR 解码步骤中提取 30 个最佳假设后，LLM（在文档级语料库上微调）选择置信度最高的假设。它们使用了 Llama-2 (Touvron et al., 2023b) 和 Mistral-7B (Jiang et al., 2023) LLM 进行翻译，且使用 Mistral-7B 进行重新排序。他们报告说，MBR 增强解码比基于 LLM 的重新排序更有效。

Sung et al. (2024) 提出了使用对话总结来有效管理上下文长度的方法。具体而言，他们使用 GPT-4o 迷你模型来生成对话总结。生成的总结以及最近的两个源-目标对被用作上下文，用 LoRA (Hu et al., 2022) 微调 Gemma (Team et al., 2024b) LLM。

Yang et al. (2024b) 使用滑动窗口方法来提高聊天翻译的性能。第一句将不带上下文翻译，并作为上下文存储以供进一步翻译。后续的句子基于缓存的源句子进行翻译。如果缓存窗口大小超过限制，最早的源句子将被移除，并存储当前输入。Llama-3 (Grattafiori et al., 2024) 基于滑动窗口方法进行微调。他们在六对语言上进行实验，使用了三种不同的窗口大小（1-3），报告指出窗口大小为 2 是最优的。他们还注意

到，使用多语言数据进行微调比使用双语数据的效果更好。

Pombal et al. (2024) 进行微调和质量感知解码 (Fernandes et al., 2022; Freitag et al., 2022) 策略的实验，以提高聊天翻译性能。在微调过程中，模型在聊天数据上进行微调，并且使用所有之前的源目标对作为上下文。对于质量感知解码，他们在 100 个候选翻译的样本上采用了（MBR）解码 (Eikema and Aziz, 2020)，并将 COMET (Rei et al., 2022) 和 Context-COMET (Agrawal et al., 2024) 作为效用函数。他们使用了 Tower LLM (Alves et al., 2024)，并报告说 MBR 解码结合微调比 ZSP 或 ICL 设置能获得更好的结果。

Wu et al. (2024d) 使用了持续预训练、有监督的微调以及对比偏好优化（CPO）(Xu et al., 2024) 来提高大型语言模型的长文本翻译性能。他们在 Llama-2 (Touvron et al., 2023b) 上进行了实验，并观察到 CPO 在与有监督的微调结合时能够提高翻译效果。他们还报告说，使用 COMET 的 MBR 解码能够获得最佳结果。

Elshin et al. (2024) 探索了深度提示调优 (Li and Liang, 2021; Liu et al., 2021) 和 CPO (Xu et al., 2024) 对视频字幕数据进行段落级翻译。它们遵循课程学习，先用句子级数据对大型语言模型进行调优，然后在训练结束时用段落级示例进行调优。它们报告说，基于课程学习的模型结合 CPO 可以获得最佳结果。

Zafar et al. (2024) 使用句子转换器 (Reimers and Gurevych, 2019) 提取了 3 个示例，并用它们对 Llama-3 (Grattafiori et al., 2024) 进行了微调以用于聊天翻译。然而，非基于 LLM 的模型（例如 NLLB (Costa-Jussà et al., 2022)）的结果在微调的 Llama-3 之上取得了最佳结果。

Luo et al. (2024) 微调了 Chinese-Llama-2 (Cui et al., 2024b) 用于中英文学翻译。他们的微调框架包括使用单语数据的持续预训练，连接的源-目标句子对，以及使用上下文感知语料库 (Guo et al., 2024) 的有监督微调。在推理过程中，所有先前翻译的句子和与源句子相似的句子都被用来保持一致性和风格相关信息。

4 其他应用

我们描述了与上下文感知翻译有关的其他杂项工作，例如用于上下文感知翻译的后编辑，使用 LLMs 评估性能等。表 3 显示了这些杂项工作的概览。

Petrick et al. (2023) 提出结合句子级别的机器翻译系统和只用文档级别单语数据训练的文档级别语言模型，以提高上下文感知的翻译性能。在解码阶段，有三种方法来估计目标词的概率：(i) 通过句子级别的机器翻译模型，(ii) 通过单语的文档级别语言模型，以及 (iii) 通过机器翻译模型学习的内部语言模型 (Herold et al., 2023)。目标词的概率通过结合这三种概率来有效估计。实验主要在一个小型语言模型（3500 万参数，在 NewsCrawl 语料库上训练 [5]）上进行，研究者报告说这种方法通常能在不增加大量计算开销的情况下得到更好的结果。他们还指出，使用 Llama (Touvron et al., 2023a)（130 亿参数）代替小型语言模型能够表现得更好。

现有的指标（如 BLEU, ChrF）不适合用于评估上下文感知翻译。Sun et al. (2025) 遵循“LLM 作为评判”(Zheng et al., 2023) 的范式来评估上下文感知翻译的质量。他们认为理想的上下文感知评估指标应该是：(i) 上下文感知的（捕捉文档级别的连贯性和准确性），(ii) 结构化的（分别评估流利性、准确性和连贯性），以及 (iii) 可解释的（易于理解并清晰识别错误）。为此，他们提出基于四个指标来评估翻译：(i) 流利性（从 1 到 5 打分），(ii) 内容错误（误译、遗漏和附加），(iii) 词汇衔接错误（不正确的词汇使用、缺少同义词和词语的过度使用），以及 (iv) 语法衔接错误（代词、连词、句子连接结构错误）。他们使用 GPT-4 作为裁判，基于上述四个指标评估翻译。他们使用调整了指令的 LLM，如 Vicuna (Chiang et al., 2023) 和 Mistral-7B (Jiang et al., 2023) 生成翻译。在两种设定下生成翻译，分别是将之前 k（范围为 1-3）个句子连接起来进行翻译，以及在单次操作中翻译整个文档。他们报告说，连接句子的翻译在 BLEU (Liu et al., 2020) 分数上相对于单次操作翻译整个文档取得了最佳结果（k=3）。然而，在针对所提的上下文感知指标方面，单次操作翻译整个文档取得了最佳结果。这表明现有的自动化指标不适合用于评估上下文感知翻译。

Mohammed and Niculae (2024) 调查了 LLMs 对正确上下文的敏感性以及它们在上下文感知翻译过程中利用上下文的效果。他们采用了两种方法来分析上下文利用：(i). 扰动分析：通过检查翻译质量和在不同类型上下文中的代词解析性能来研究模型的稳健性。(ii). 归因分析：通过 ALTI-Logit (计算源词对特定目标词的贡献) (Ferrando et al., 2023) 和输入擦除法 (衡量移除部分输入后模型预测的变化) (Li et al., 2016) 的归因方法，分析翻译过程中上下文相关部分的贡献。他们在九种不同的 LLMs 上进行了实验，涉及英译法和英译德语言对。在扰动分析中，他们对三种不同类型的上下文

[5] https://data.statmt.org/news-crawl/

Work	LLMs	Language Pairs
Automatic Post-Editing		
Koneru et al. (2024)	Llama-2	En → De
Li et al. (2025)	Llama-3	En → De
Dong et al. (2025)	Llama-3, Mistral-Nemo-Instruct	En ↔ { Fr, De, Ru, Es, Zh }
Agentic Framework		
Wang et al. (2024b)	GPT-3.5-Turbo, GPT-4o-mini, Qwen2	En ↔ { Fr, De, Zh, Ja }
Briakou et al. (2024)	Gemini-1.5-Pro	En → { De, Cs, Ru, Ja, He, Uk, Zh }
Wu et al. (2024c)	GPT-4	Zh → En
Guo et al. (2025)	Qwen2.5, Llama-3.1	En ↔ { Fr, De, Zh, Ja }
Miscellaneous Works		
Petrick et al. (2023) (Combining sentence-level NMT with document-level LM)	Llama	En → { De, It }
Sun et al. (2025) (LLM-as-Judge)	GPT-4, Vicuna, Mistral	En ↔ { De, Zh }
Mohammed and Niculae (2024) (Analysis of context usage)	EuroLLM, Llama-2, ALMA, TowerBase, TowerInstruct	En → { Fr, De }

Table 3: 本文涵盖的 APE、代理框架及其他各种方法列表。

进行了实验：(i). Gold（作为上下文的前一个源-目标对），(ii). Perturbed（随机从其他文档中抽取的句对），(iii). Random（从模型词汇表中均匀抽取的随机词）并利用对比测试集 (Müller et al., 2018; Lopes et al., 2020; Post and Junczys-Dowmunt, 2023) 进行代词解析实验。他们采用 ICL，并以五个先前的源-目标对为上下文。所有模型对随机上下文是稳健的（就自动度量而言），这表明适当上下文利用的缺乏。即便如此，在对比测试集中随机和扰动上下文的效果更为明显。他们报告指出需要进行明确的上下文感知微调以更好地利用上下文。

Wang et al. (2024b) 提出了一种基于文档的翻译代理，该代理具有四个记忆组件。这些记忆组件包括：(i) 专有名词记录：其中包含以前翻译的源-目标名词对。当识别出新的专有名词时，专有名词记录会更新，如果该名词已存在于记忆中，则会使用其对应的翻译。这是为了在整个文档中保持翻译的一致性。(ii) 双语摘要：包含源语言和目标语言的摘要。摘要组件确保连贯性，并有助于产生更一致的翻译。(iii) 长期记忆 (iv) 短期记忆：长期和短期记忆都包含以前翻译的源-目标句子对。具体来说，短期记忆旨在捕捉当前句子周围的立即上下文，长期记忆则设计用于捕捉更广泛的上下文。这四个记忆组件的信息被整合到一个用于逐句翻译文档的提示中，确保源文档中的所有句子在保持一致性和流畅性的同时得以翻译。他们在 GPT-3.5、GPT-4 (Brown et al., 2020; Achiam et al., 2023) 和 Qwen (Bai et al., 2023) LLMs 上进行了实验。

Briakou et al. (2024) 提出了一个文档翻译的链式思维 (Wei et al., 2022) 方法。该框架由四个阶段组成，在此期间，对源文档进行翻译和改进，以提高翻译的准确性和流畅性。他们在 Gemini 1.5 Pro (Team et al., 2024a) LLM 上进行了实验，所提出的方法在翻译性能上优于 ZSP 方法。

Wu et al. (2024c) 提出了一个用于文学翻译的多智能体虚拟代理框架。他们的方法模仿了翻译工作台中的各种层级（例如，翻译、校对、编辑等）。他们的框架可以有效地翻译非常长的文学文本，并在人类评估指标中更受欢迎，但无法翻译较短的文本。

使用大型语言模型（LLMs）进行整篇文档翻译（文档到文档翻译）时的主要问题之一是源文档中的某些句子未被翻译。为了翻译给定文档中的每个源句子，句子需要逐句处理，这常常导致一致性和流畅性问题（如缺乏上下文连贯性）。为了在没有一致性问题的情况下确保每个句子的翻译，Guo et al. (2025) 引入了一种基于提示的代理，采用增量句子级强制解码策

略。具体来说，文档是逐句翻译的，以确保所有源句子都被翻译。然而，输入给 LLM 的是两个源句子（即，$s_{(i-1)}$ 和 s_i），而之前的源句子的翻译结果（即，$t_{(i-1)}$）作为强制解码段。这迫使 LLM 再生成之前翻译的句子，并在翻译整个文档的过程中保持流畅。最后，将当前源句子（即，t_i）的翻译拼接以形成最终翻译。他们还通过将源文档和目标文档的摘要以记忆的形式进行了增强，以提高话语翻译。他们在 Qwen (Bai et al., 2023) 和 Llama-3 (Grattafiori et al., 2024) LLMs 上进行了实验。他们发现，当强制解码的前文句子超过一个时，性能上没有显著差异。

自动后编辑（APE）(Pal et al., 2016) 是一种提高任何机器翻译系统翻译性能的有效方法。一些研究利用大型语言模型（LLMs）探索了文档级翻译的 APE。Koneru et al. (2024) 提出了使用 LLM 来后编辑来自神经机器翻译（NMT）模型的翻译。LLM 在（源文本、假设、参考）三元组上进行微调，其中假设是由 NMT 模型生成的。他们为句子和文档级输入都训练了 APE 模型。具体来说，文档级句子是逐句独立翻译的，然后连接在一起作为输入提供给 APE 模型。他们使用 Llama-2 (Touvron et al., 2023b)，遵循 LoRA (Hu et al., 2022) 来训练 APE 模型。他们报告称，APE 是必要的，并显著改善了在自动和上下文感知评估指标上相较于 ICL 方法和微调的 LLM 的翻译性能。通过使用黄金上下文而非使用模型生成的输出进行文档级机器翻译，可以进一步提高 APE 的性能。

与 Koneru et al. (2024) 类似，Li et al. (2025) 也利用 LLM 对来自非 LLM 为基础的句子级 MT 模型的翻译进行后编辑。他们通过使用目标到源 NMT 模型翻译单语文档级目标数据来创建合成数据以训练 LLM，然后利用创建的合成源进一步通过源到目标 NMT 模型生成合成的目标数据。此过程旨在生成合成的自然目标数据，作为监督训练数据来训练 LLM。他们通过两个目标对 LLM 进行微调，一个仅使用合成且干净的数据，另一个使用合成目标到源和干净目标数据。尽管使用源和目标数据来训练 APE 模型是直观的，但他们观察到这会导致性能较差。其原因在于 APE 模型是用合成的源数据训练的，但在推理期间，干净的源数据连同 NMT 模型的输出（合成目标）一起被输入模型。他们进一步提出两种方法来解决这个问题：(i) 级联方法，其中 LLM 微调以对合成源数据进行 APE，结果输出与合成目标数据一起用于对合成目标数据进行 APE；(ii) 连续预训练方法，其中 LLM 最初用合成和自然的源和目标数据进行预训练，然后用合成的源到目标和干净的目标数据进行微调。他们在 Llama-3 (Grattafiori et al., 2024) 上进行了实验，结果表明所提出的 APE 模型在自动和上下文感知评估指标方面，性能优于仅进行翻译训练的 Llama 模型。有趣的是，仅使用目标数据训练的 APE 模型与同时使用源-目标数据（即级联和连续预训练方法）训练的模型在性能上只有细微的差别。同样地，Thai et al. (2022) 对 GPT-3 进行了微调，以对 Google 翻译生成的机器翻译输出进行后编辑。

Dong et al. (2025) 微调的 Llama-3 和 Mistral-Nemo-Instruct [6] 用于后编辑文档级别的翻译。起初，他们生成了两种翻译集（句子级和文档级）并用它们来训练一个 APE 模型。他们观察到句子级别的翻译减少了幻觉，而文档级别的翻译减少了不一致性。因此，使用这两种设置的翻译可以提高文档级别翻译的改进过程。

我们对使用大型语言模型（LLM）来进行上下文感知机器翻译的以往研究进行了文献综述。基于提示的方法（零样本和少量样本提示）表明，LLM 可以有效地处理上下文感知翻译任务，并且利用可用的语料进行微调可在多种语言对（如英语、法语、德语、西班牙语和汉语）中达到先进水平的表现。在基于提示的方法中，闭源的 LLM（如 ChatGPT 和 Tower LLM）的表现优于开源的 LLM（如 Llama 和 Bloom）。然而，对 Llama 或 Bloom LLM 进行微调可以达到与 ChatGPT 或 Tower LLM 几乎相同的性能。LLM 还被用于自动后编辑任务，以后编辑翻译内容，并在捕捉篇章级现象方面表现良好。我们观察到，使用 LLM 进行上下文感知翻译是一个相对较新的研究方向，已有的工作较少。为此，我们建议以下未来的研究方向：

面向上下文的低资源语言翻译： 不同于句子级语料库，许多语言对没有可用的文档级语料库，但大量的单语文档级语料库是可用的。可以使用大型语言模型为缺乏或没有文档级语料库的语言对，通过提示或自动后编辑来构建具备上下文意识的翻译系统。

翻译代理： 代理框架 Wu et al. (2024c); Briakou et al. (2024) 是构建情境感知机器翻译系统的另一个有趣的研究方向。翻译的不同方面可以由各种代理处理（如词汇翻译一致性、整体流畅度、情境感知词语和短语的跟踪等），类似于 Wang et al. (2024b)。

上下文感知评估： 健壮且可解释的评估指标对于评估机器翻译输出质量至关重要。LLMs 可以显著影响 (Agrawal et al., 2024) 对 MT 系统的评估。未来的方向可以探索使用 LLMs 及

[6] https://mistral.ai/news/mistral-nemo

最少的监督数据来有效地评估任意语言对的翻译质量。

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 31–40, Alicante, Spain.

Sweta Agrawal, Amin Farajian, Patrick Fernandes, Ricardo Rei, and André F. T. Martins. 2024. Assessing the role of context in chat translation evaluation: Is context helpful and under what conditions? *Transactions of the Association for Computational Linguistics*, 12:1250–1267.

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. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Ramakrishna Appicharla, Baban Gain, Santanu Pal, and Asif Ekbal. 2023. A case study on context encoding in multi-encoder based document-level neural machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 160–172, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Ramakrishna Appicharla, Baban Gain, Santanu Pal, Asif Ekbal, and Pushpak Bhattacharyya. 2024. A case study on context-aware neural machine translation with multi-task learning. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 246–257, Sheffield, UK. European Association for Machine Translation (EAMT).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. 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 3442–3455, Online. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

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, Miami, Florida, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024a. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association*

for Computational Linguistics: ACL 2024, pages 10885–10897, Bangkok, Thailand. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024b. Efficient and effective text encoding for chinese llama and alpaca. *Preprint*, arXiv:2304.08177.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Domenic Donato, Lei Yu, and Chris Dyer. 2021. Diverse pretrained context encodings improve document translation. 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 1299–1311, Online. Association for Computational Linguistics.

Yichen Dong, Xinglin Lyu, Junhui Li, Daimeng Wei, Min Zhang, Shimin Tao, and Hao Yang. 2025. Two intermediate translations are better than one: Fine-tuning llms for document-level translation refinement. *Preprint*, arXiv:2504.05614.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, Dmitry Popov, Anton Chekashev, Vladislav Negodin, Vera Frantsuzova, Alexander Chernyshev, and Kirill Denisov. 2024. From general LLM to translation: How we dramatically improve translation quality using human evaluation data for LLM finetuning. In *Proceedings of the Ninth Conference on Machine Translation*, pages 247–252, Miami, Florida, USA. Association for Computational Linguistics.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *arXiv preprint arXiv:2404.13813*.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Javier Ferrando, Gerard I Gállego, Ioannis Tsiamas, and Marta R Costa-jussà. 2023. Explaining how transformers use context to build predictions. *arXiv preprint arXiv:2305.12535*.

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.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2025. Bridging the linguistic divide: A survey on leveraging large language models for machine translation. *Preprint*, arXiv:2504.01919.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiaxin Guo, Yuanchang Luo, Daimeng Wei, Ling Zhang, Zongyao Li, Hengchao Shang, Zhiqiang Rao, Shaojun Li, Jinlong Yang, Zhanglin Wu, and 1 others. 2025. Doc-guided sent2sent++: A sent2sent++ agent with doc-guided memory for document-level machine translation. *arXiv preprint arXiv:2501.08523*.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 639–649, Mexico City, Mexico. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Christian Herold, Yingbo Gao, Mohammad Zeineldeen, and Hermann Ney. 2023. Improving language model integration for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7114–7123, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hanxu Hu, Jannis Vamvas, and Rico Sennrich. 2025. Source-primed multi-turn conversation helps large language models translate documents. *arXiv preprint arXiv:2503.10494*.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.

Keito Kudo, Hiroyuki Deguchi, Makoto Morishita, Ryo Fujii, Takumi Ito, Shintaro Ozaki, Koki Natsumi, Kai Sato, Kazuki Yano, Ryosuke Takahashi, Subaru Kimura, Tomomasa Hara, Yusuke Sakai, and Jun Suzuki. 2024. Document-level translation with LLM reranking: Team-J at WMT 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 210–226, Miami, Florida, USA. Association for Computational Linguistics.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yachao Li, Junhui Li, Jing Jiang, Shimin Tao, Hao Yang, and Min Zhang. 2023. P-transformer: Towards better document-to-document neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3859–3870.

Yachao Li, Junhui Li, Jing Jiang, and Min Zhang. 2024. Enhancing document-level translation of large language model via translation mixed-instructions. *arXiv preprint arXiv:2401.08088*.

Zongyao Li, Zhiqiang Rao, Hengchao Shang, Jiaxin Guo, Shaojun Li, Daimeng Wei, and Hao Yang. 2025. Enhancing large language models for document-level translation post-editing using monolingual data. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8830–8840, Abu Dhabi, UAE. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and 1 others. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Yuanchang Luo, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhiqiang Rao, Shaojun Li, Jinlong Yang, and Hao Yang. 2024.

Context-aware and style-related incremental decoding framework for discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 973–979, Miami, Florida, USA. Association for Computational Linguistics.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024a. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Xinglin Lyu, Junhui Li, Yanqing Zhao, Min Zhang, Daimeng Wei, Shimin Tao, Hao Yang, and Min Zhang. 2024b. DeMPT: Decoding-enhanced multi-phase prompt tuning for making LLMs be better context-aware translators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20280–20295, Miami, Florida, USA. Association for Computational Linguistics.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Wafaa Mohammed and Vlad Niculae. 2024. Analyzing context utilization of llms in document-level translation. *arXiv preprint arXiv:2410.14391*.

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, Singapore. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Berlin, Germany. Association for Computational Linguistics.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Ziqian Peng, Rachel Bawden, and François Yvon. 2024. *Handling Very Long Contexts in Neural Machine Translation: a Survey*. Ph.D. thesis, Projet ANR MaTOS.

Frithjof Petrick, Christian Herold, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2023. Document-level language models for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 375–391, Singapore. Association for Computational Linguistics.

Jose Pombal, Sweta Agrawal, and André Martins. 2024. Improving context usage for translating bilingual customer support chat with large language models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 993–1003, Miami, Florida, USA. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *arXiv preprint arXiv:2304.12959*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 173–185, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Yirong Sun, Dawei Zhu, Yanjun Chen, Erjia Xiao, Xinghao Chen, and Xiaoyu Shen. 2025. Fine-grained and multi-dimensional metrics for document-level machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 1–17, Albuquerque, USA. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Mingi Sung, Seungmin Lee, Jiwon Kim, and Sejoon Kim. 2024. Context-aware LLM translation system using conversation summarization and dialogue history. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1011–1015, Miami, Florida, USA. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024a. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.

Longyue Wang, Siyou Liu, Mingzhou Xu, Linfeng Song, Shuming Shi, and Zhaopeng Tu. 2023a. A survey on zero pronoun translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3325–3339, Toronto, Canada. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023b. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2024b. Delta: An online document-level translation agent based on multi-level memory. *arXiv preprint arXiv:2410.08143*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024b. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024c. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *arXiv preprint arXiv:2405.11804*.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169, Singapore. Association for Computational Linguistics.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Hengchao Shang, Jiaxin Guo, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Ning Xie, and Hao Yang. 2024d. Choose the final translation from NMT and LLM hypotheses using MBR decoding: HW-TSC's submission to the WMT24 general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 155–164, Miami, Florida, USA. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *Preprint*, arXiv:2401.08417.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xinye Yang, Yida Mu, Kalina Bontcheva, and Xingyi Song. 2024b. Optimising LLM-driven machine translation with context-aware sliding windows. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1004–1010, Miami, Florida, USA. Association for Computational Linguistics.

Maria Zafar, Antonio Castaldo, Prashanth Nayak, Rejwanul Haque, and Andy Way. 2024. The SETU-ADAPT submissions to WMT 2024 chat translation tasks. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1023–1030, Miami, Florida, USA. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, and 1 others. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022a. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4176–4192, Dublin, Ireland. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. 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, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.