从话语到生动表达：通过自适应局部偏好优化训练富有表现力的字幕翻译大语言模型

崔超群[1,2], **王世静**[3], **黄梁斌**[3],**顾青青**[4],**黄兆龙**[3],**曾晓**[3], **毛文吉**[1, 2, *]中国科学院自动化研究所 MAIS[2]中国科学院大学人工智能学院[3]沪京数字媒体与娱乐集团[4]吉利人工智能实验室 cuichaoqun2025@ia.ac.cn, wenji.mao@ia.ac.cn

arXiv:2602.01068v1 [cs.CL] 1 Feb 2026

摘要

大语言模型（LLM）的快速发展显著提升了机器翻译的通用能力。然而，随着应用场景日益复杂，大语言模型在垂直领域翻译中的局限性逐渐显现。本研究聚焦于如何构建满足领域定制需求的翻译大语言模型。我们以视听媒体字幕翻译为主题，探索如何训练富有表现力、生动鲜活的翻译大语言模型。我们考察了字幕翻译以及其他领域直译与意译的情况，并验证了将大语言模型用作翻译奖励模型和评估器的可靠性。此外，为了训练富有表现力的翻译大语言模型，我们构建并发布了一个多方向字幕平行语料库数据集，并提出自适应局部偏好优化（ALPO）方法，以解决细粒度偏好对齐问题。实验结果表明，ALPO 在翻译质量的多维度评估中表现出色。

1 引言

近年来，大语言模型（LLM）的快速发展显著提升了机器翻译的通用能力（Xu et al., 2024c; Enis & Hopkins, 2024; Feng et al., 2024）。以 GPT-4（Achiam et al., 2023）和通义千问 Qwen3（Yang et al., 2025）为代表的大语言模型经过海量跨语言数据训练，在多语言翻译任务中展现出卓越的上下文理解与生成能力。它们在通用领域（如新闻、日常对话）的翻译质量已接近人工水平（Xu et al., 2024a;b）。然而，随着应用深入垂直领域，大语言模型在专业翻译场景中的局限性日益凸显：专业术语一致性不足（An et al., 2024; Huang et al., 2024）、偏离行业标准表达（Ling et al., 2023; Pang et al., 2025）以及风格适应性较弱（Zhang et al., 2024）等问题尤为突出。因此，基于大语言模型的机器翻译研究日益聚焦于满足领域特定定制需求，例如开发严格遵循条款起草规范的法律法规领域大语言模型，或精准匹配医学术语的医学领域大语言模型。

在本研究中，我们聚焦于视听媒体字幕翻译任务，旨在将电影、电视剧和纪录片等不同类型视听媒体节目中的字幕台词从源语言翻译成目标语言。这项任务在娱乐节目本地化以及促进全球文化传播与交流方面发挥着至关重要的作用（Federico 等人，2020；Wu 等人，2023；Mhaskar 等人，2024），但它仍然是机器翻译中尚未得到充分探索的子领域。与文学翻译类似，字幕翻译要求对节目台词进行本地化且具有风格性的**意译**，以传达原台词的氛围、情感和语气。

* 通讯作者。

然而，尽管大语言模型能够实现较高的翻译准确性，却往往倾向于**直译**（正如我们在第 3.2 节中的定量分析所示）。因此，探索如何训练出兼具较强表现力和生动性的定制字幕翻译大语言模型，是我们面临的首要挑战。

我们采用 LLM-as-a-Judge（使用大语言模型作为评估器）和偏好优化技术，构建定制字幕翻译大语言模型。我们定量考察了不同领域中人工翻译文本的意译程度，以及各类大语言模型在字幕翻译中的意译表现。研究结果表明：（1）与法律文本、新闻和医学等要求高准确性的严肃文本翻译（直译）相比，字幕翻译更倾向于意译；（2）在字幕翻译中，聊天大语言模型生成的译文相比推理大语言模型和人类产出的译文，往往更加直译。我们还通过实验验证了大语言模型作为字幕翻译奖励模型和评估器的可靠性，以及它们与人类偏好的对齐程度。这使得自动构建用于偏好对齐训练的偏好数据成为可能，从而提升字幕翻译大语言模型的表现力和生动性。此外，针对短字幕行翻译任务——该任务需要细粒度的局部偏好对齐——我们提出了一种新颖的偏好对齐策略，称为**自**适应**局**部**偏**好**优**化（ALPO）。

综上所述，本研究的主要贡献如下：

- 我们首次提出视觉媒体字幕翻译任务，并探究该领域中意译的程度以及 LLM-as-a-Judge 的可靠性。
- 我们提出 ALPO，一种局部偏好优化方法，对其有效性给出形式化解释，并利用该方法训练出具有高生动性的字幕翻译大语言模型。
- 我们发布了多个方向的字幕平行语料库以支持社区研究，并建立了基于 LLM-as-a-Judge 的多维评估框架。
- 实验结果表明，我们使用 ALPO 训练的140亿参数大语言模型在多个维度上取得了显著提升，并优于最先进的大语言模型。

2 相关工作

2.1 LLM-as-a-Judge

大语言模型模拟人类推理，并依据一组预定义规则评估特定输入的能力，为“LLM-as-a-Judge”铺平了道路（Zheng 等，2023a；Gu 等，2024）。现有研究表明，大语言模型具备可扩展性、适应性和成本效益，因而特别适合处理日益增多、传统上由人类完成的评估任务（Chen 等，2024a；Li 等，2024a；Wang 等，2024c；Zhu 等，2025）。这些能力对于在多样化的评估场景和目标中灵活部署大语言模型至关重要，也推动了它们在实际评估场景中的迅速普及（Wang 等，2024b；Chen 等，2024b；Khattak 等，2023）。

大语言模型最初用于语言生成和理解任务，后来通过基于人类反馈的强化学习（RLHF）等先进训练方法（Ouyang 等，2022）取得了显著发展，从而增强了其与人类价值观及推理过程的对齐。这种对齐使大语言模型得以从生成任务转向评估角色。从根本上说，LLM-as-a-Judge 是指部署大语言模型，依据预定义规则、标准或偏好来评估对象、行为或决策（康德，1908；1987）。这一框架涵盖了多种评估角色，包括：评分器（Dong 等，2023；Luong 等，2024）、评估器（Li 等，2024b；Zhang 等，2023）、评论器（Ke 等，2024；Putta 等，2024；Xiong 等，2024）、验证器（Shinn 等，2023；Wang 等，2024d）、审查器（Bai 等，2023）、奖励/排序模型（Luo 等，2023；Sun 等，2024；Yang 等，2024b）等。

2.2 语言模型偏好优化

强化学习为使大语言模型与人类价值观对齐并控制文本生成提供了有效方案（Bender 等人, 2021; Bommasani 等人, 2021; Thoppilan 等人, 2022; Taori 等人, 2023; Chiang 等人, 2023; Ji 等人, 2023）。为此，研究者建立了基于人类反馈奖励模型的基于人类反馈的强化学习框架（Ouyang 等人, 2022; Christiano等人, 2017;

MacGlashan 等人, 2017; Ziegler 等人, 2019; Stiennon 等人, 2020; Zheng 等，2023b）。然而，尽管基于人类反馈的强化学习卓有成效，其复杂性、不稳定性以及对超参数的敏感性仍未得到充分解决（Andrychowicz 等人, 2021; Engstrom 等人, 2020）。近期提出的直接偏好优化（DPO）（Rafailov 等，2024年）通过免去显式构建奖励模型或执行强化学习程序的需要，简化了基于人类反馈的强化学习框架，从而避免了对奖励模型的依赖。此后又出现了若干变体，例如 SimPO、KTO 和 IPO（Meng 等人, 2024; Azar 等人, 2024; Ethayarajh 等人, 2024）。然而，将这些方法用于局部偏好对齐任务时，仍存在一些局限，包括粒度较粗和梯度稀释（形式化定义见附录 D）。

3 实证研究

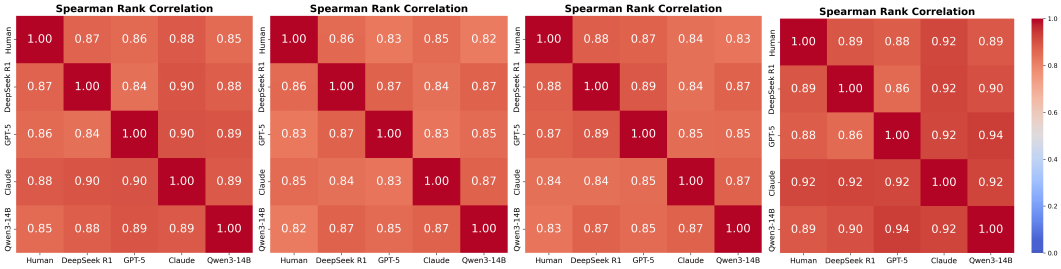
3.1 大语言模型是优秀的翻译评估器

我们旨在验证，大语言模型能否在字幕翻译各个维度的质量评估中与人类偏好高度一致，从而使大语言模型能够同时作为字幕翻译模型对齐训练的奖励模型和评估器。我们首先对大语言模型评估器进行形式化定义：

$$\mathcal{E} \leftarrow \pi_{\mathrm{e}}(\mathcal{C} \oplus s \oplus \{t\}), \tag{1}$$

其中， π_{e} 是用于评估的大语言模型， $\mathcal{C}$ 是提示中的介绍和指令，用于指导大语言模型进行评估； s 是待评估的句子； $\{t\}$ 表示 s的多个译文； $\mathcal{E}$ 表示 π_{e}给出的评估分数。

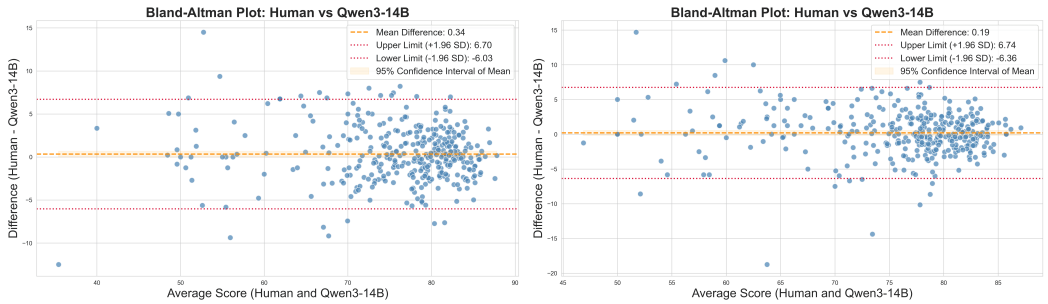
我们研究了人工评估者与大语言模型评估器在生动性评分方面的相关性。研究对象是附录 A 中自建的**Mu**多语言**S**字幕**C**语料库（MuSC）数据集中 500 个句子的 10种不同译文。每位评估者都对每个待评估句子的 10 个译文进行 0 到 100 分的评分，随后我们计算了评估者之间的平均斯皮尔曼等级相关系数 ρ。图 1 展示了斯皮尔曼等级相关系数的结果。详细的实验设置见附录 B.2.1。翻译方向包括 en⇒zh、en⇒de、zh⇒en 和 zh⇒th，其中中文和英语、德语、泰语分别对应高、中、低资源语言。



(a) en⇒zh (b) en⇒de (c) zh⇒en (d) zh⇒th

图1：利用斯皮尔曼等级相关系数 ρ考察多个评估器的一致性。

结果表明，LLM 评估器在多个方向上与人工评估者表现出高度一致，说明它能够处理不同语言间文化和语言惯例的差异。值得注意的是，即使是拥有 14B 参数的 Qwen3-14B 模型，在所有方向上与人工评估者及其他最先进的大语言模型也都达成了高度一致（$\rho \geq 0.82$）。这表明，推理成本相对较低的 14B 模型可以充当高效的奖励模型，为 ALPO 奠定基础。此外，图 2 展示了 Qwen3-14B 与人工评估者对 en⇒zh 和 zh⇒th 中相同字幕译文所作评分的 Bland-Altman 图。结果表明，两者的平均差（MD）非常小，系统性偏差极低，且一致性界限（LoA）处于 100 分制评分可接受的误差范围内。这进一步证实了 14B 模型与人工评估者之间的一致性。



(a) en⇒zh (b) zh⇒th

图 2：通义千问Qwen3-14B与人工评估者的 Bland-Altman图。

3.2 并行语料 行语料实际上并不平行 ALLEL

直译侧重翻译准确性，力求保留原文形式；而意译强调传达含义，并灵活调整表达方式（Baker与Saldanha，2019年；Munday等人，2022）。根据我们的经验和观察，不同领域的译文所采用的意译程度各不相同：字幕和文学翻译等领域通常倾向于意译，而法律文本、医学和技术文本通常要求高度准确的直译。

我们设计了相关实验，以定量验证这一假设。我们考察了不同领域平行语料库的回译一致性。具体而言，对于特定的翻译方向，我们使用 GPT-4o 将目标语言译文直接回译为源语言，然后计算回译文本与原文之间的 BLEU 和 ChrF++ 分数。分数越低，表示意译程度越高。详细实验设置见附录 B.2.2。

表 1 的调查结果表明，与法律文本、新闻和医学领域相比，视觉媒体、文学和宗教等领域的回译文本与原文的相似度较低。这表明其意译程度更高。对于视听媒体字幕翻译，译者会将视频信息融入译文，力求传达原文的氛围、情感和语气；这使译文更具表现力和生动性，有时会以牺牲一定程度的准确性为代价。

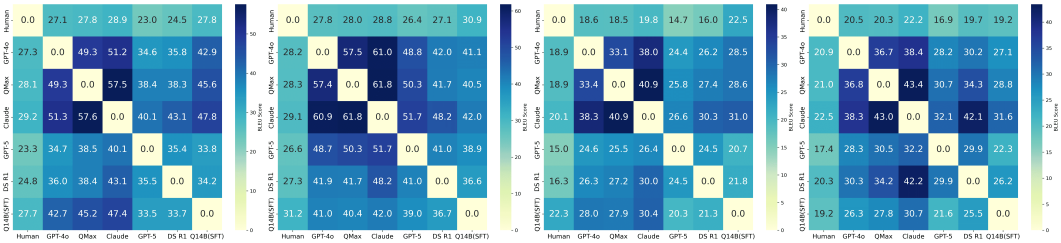
3.3 聊天大语言模型偏好直译，推理大语言模型擅长意译

一般而言，能够有效进行意译的大语言模型往往会生成更具表现力的译文。我们旨在验证不同大语言模型的意译能力。我们的研究涵盖多个翻译方向，每个方向均从 MuSC测试集抽取 2,000 行，并提示多种大语言模型生成富有表现力、生动鲜活的译文（提示见附录 F.1）。随后，我们计算了人工译文与多个大语言模型生成的译文之间的两两 BLEU 相似度。纳入研究的大语言模型包括 GPT-4o（Achiam 等，2023）、通义千问-Max（Yang 等，2024a）和 Claude Opus 4.1（Anthropic，2025）等聊天模型，以及 GPT-5 Thinking（OpenAI，2025）和 DeepSeek-R1（Guo 等，2025）等推理模型。此外，我们还纳入了在 MuSC 上训练的监督微调（SFT）通义千问2.5-14B模型。

图 3 的结果表明，聊天模型生成的译文彼此之间相似度较高，说明它们倾向于采用更直译的方式。相比之下，推理模型生成的译文与其他译文的相似度较低，表明这些模型通过思考，能够更好地遵循意译指令。这验证了大语言模型的推理时扩展能够有效提升翻译性能。人工翻译与其他译文的相似度较低，表明其擅长意译。我们在附录 C.2 中展示了不同译文的示例，以直观呈现这些发现。

Ta表 1：领域意译研究。最高值和最低值分别以蓝色和橙色表示 .

数据集	领域	en⇒de		en⇒fr		en⇒es	
		BLEU (					
OpenSubtitles	视觉媒体	**15.00**	**37.08**	**17.84**	**37.80**	21.60	43.78
书籍	文学	17.22	43.20	21.51	47.14	**18.60**	**43.00**
bible-uedin	宗教	22.55	49.57	22.07	48.24	22.31	51.03
DGT	法律文本	19.85	50.05	22.44	48.10	21.88	52.24
JRC-Acquis		**27.55**	**61.95**	**28.83**	**69.65**	25.16	65.30
News-Commentary	News	25.00	55.23	24.87	54.37	**35.90**	65.84
ECDC	医学	23.18	60.77	23.86	62.97	27.20	65.39
EMEA		25.20	59.36	24.73	61.65	31.46	**69.15**
数据集	**领域**	德语⇒英语		法语⇒英语		西班牙语⇒英语	
		BLEU (					
OpenSubtitles	视觉媒体	15.25	**39.41**	**16.60**	**40.65**	22.18	47.45
书籍	文学	**14.50**	40.50	20.12	48.29	**15.84**	**43.99**
bible-uedin	宗教	15.86	47.14	22.79	51.85	24.55	53.77
DGT	法律文本	19.70	47.77	23.31	48.91	22.87	52.49
JRC-Acquis		**27.54**	**63.43**	**31.33**	64.15	27.68	64.46
News-Commentary	News	22.64	54.05	29.34	56.20	**41.75**	66.35
ECDC	医学	21.09	57.41	29.13	**64.81**	29.62	64.20
EMEA		24.94	60.16	26.24	62.96	29.86	**67.09**



(a) en⇒zh (b) en⇒de (c) zh⇒en (d) zh⇒th

图 3：译文之间的两两 BLEU 分数。

4 方法

4.1 符号说明

用于字幕翻译的双语数据集 $\mathbb{D}$ 由多个不同类型项目（如电影或电视剧）中的源语言语句 $\mathcal{L}_{\text{src}}$ 和目标语言语句 $\mathcal{L}_{\text{tgt}}$ 构成，表示为 $\mathbb{D} \equiv \{(\ell_{\text{src}},\ \ell_{\text{tgt}}) \in \mathcal{L}_{\text{src}} \times \mathcal{L}_{\text{tgt}}\}$。同一项目的多语言字幕中的语句通常并非一一对应。为此，我们开发了一种复杂度为 $O(N)$ 的双语平行语料库构建算法；详情见附录A.2。我们将使用数据集 $\mathbb{D}$ 对现成的大语言模型进行监督微调和 ALPO 训练。大语言模型的输入提示 x 包括上下文 $\mathcal{C}$（由引言和指令组成）以及一组需要翻译的 n 源语句$\{s_i \mid i \in [n]\}$，即 $x \equiv \mathcal{C} \oplus \{s_i\}$。大语言模型的响应 y 包含每个语句的译文 $\{t_i \mid i \in [n]\}$，即 $y \equiv \{t_i\}$。详细的 x 和 y 格式见附录 F.1。

4.2 总体框架

显然，构建富有表现力的翻译大语言模型是一项偏好优化任务。然而，我们无法直接应用通用方法（例如近端策略优化（Ouyang 等，2022年）或 DPO（Rafailov 等，2024年）），主要原因是每行字幕的翻译都取决于其上下文。SFT模型的输入 x π_{sft} 必须包含多行字幕 $\{s_i\}$，因此需要对响应 y中的每一行进行更细粒度的对齐。结果监督方法（如近端策略优化和 DPO）会对大语言模型的完整输出进行优化，因此不足以胜任此任务。我们在附录

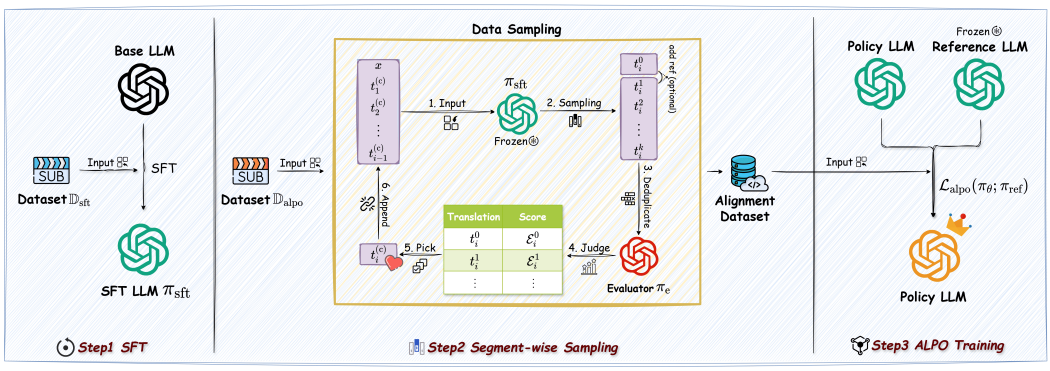


图 4：ALPO 的总体框架。

C.4 和附录 D 中对这一点进行了实验和理论验证。为解决这项细粒度偏好优化任务，我们提出了过程监督的 ALPO 方法。ALPO 引入了一种新范式，通过分段采样策略和细粒度对齐损失来训练高质量的字幕翻译大语言模型。ALPO 的总体框架如图 4 所示。

4.3 采样策略

我们将需要对大语言模型响应进行多片段局部偏好对齐的任务称为**局部偏好优化** 任务。ALPO 为处理此类任务提供了一种有效范式。我们首先将平行语料库 $\mathbb{D}$ 划分为示范数据集 $\mathbb{D}_{\text{sft}}$ 和对齐数据集 $\mathbb{D}_{\text{alpo}}$（约为 8:2），使用 $\mathbb{D}_{\text{sft}}$ 训练 SFT模型 π_{sft} （输入输出格式见附录 F.1）。

在采样阶段，对于一个 $x \in \mathbb{D}_{\text{alpo}}$，其中包含 n 行 s_1、……、s_n。对于每个 s_i，以 $p_i = x$、$t_1^{(c)}$、……、$t_{i-1}^{(c)}$ 为前缀，采样得到 k 个译文（实验中为 $k = 15$），即 $\pi_{\text{sft}}(t_i^j \mid p_i)$、$j = 1$、……、$k$，从而得到候选集 $\{t_i^j \mid j = 1, \ldots, k\}$。随后，对候选集进行去重；如果存在人工参考 t_i^0，则可将其加入候选集，形成 $\mathcal{T}_i = \{t_i^j \mid j = 0、1、\ldots\}$。接着，使用 Qwen3-14B 模型作为评估器 π_e（也可视为奖励模型），评估 $\mathcal{T}_i$的生动性，得到分数序列 $\mathcal{E}_i$（示范见表 2）。基于 $\mathcal{E}_i$，我们为 s_i 选择一个较优译文 $t_i^{(c)}$（从得分最高的 3 个译文中随机选取），作为下一采样循环的前缀。最终，对于一个 $x \in \mathbb{D}_{\text{alpo}}$，我们得到采样结果 $\mathcal{S}(x) \equiv \{(s_i、\mathcal{T}_i、\mathcal{E}_i) \mid i = 1、\ldots\ldots、n\}$。算法 1 详细介绍了整个流程。

表 2：评估器 π_e 的 zh⇒en 评估演示。

Line	对 π_e 的评估
历史虽然会重演，但是人类是无法回到过去的。	-
历史会重演，但我们无法回到曾经的样子。	70
历史或许会重演，但人类无法让时间倒流。	85
历史的车轮或许会转满一圈，但通往过去的门永远紧锁。	88
虽然历史可能重演，但人类无法回到过去。	82
历史常常回响，但我们无法让时光倒流。	92

4.4 自适应对齐损失

奖励信号作用于模型输出片段（过程监督），而非完整响应（结果监督），这构成了生动性增强的主要挑战。ALPO 能够对翻译大语言模型响应y中的多个片段进行过程监督式对齐。具体而言，对于一个 $x \in \mathbb{D}_{\text{alpo}}$中的所有行 $\{s_i \mid i = 1, \ldots, n\}$，我们为每个 s_i 分配一个自适应权重$w(s_i)$，其定义为门控函数$\mathbf{1}(s_i)$ 与重要性得分 $\delta(s_i)$ 的乘积：

$$w(s_i) = \mathbf{1}(s_i) \cdot \delta(s_i), \tag{2}$$

其中 $\mathbf{1}(s_i)$ 充当门控，用于确定 s_i 是否参与优化。当采样翻译缺乏多样性或明显的质量差异，即 $|\mathcal{T}_i| \leq 3$ 或 $\max(\mathcal{E}_i) - \min(\mathcal{E}_i) \leq 5$ 时，我们

算法 1 ALPO采样策略。

输入： SFT模型 π_{sft}、评估大语言模型 π_{e}、对齐数据集 $\mathbb{D}_{\text{alpo}}$、采样数量 k。
输出： 采样片段级候选集 $\mathcal{S}(x)$。

对于任意 $x \in \mathbb{D}_{\text{alpo}}$ **执行** //遍历对齐数据集 $\mathbb{D}_{\text{alpo}}$。
  循环 $i = 1$至 n **执行** // 遍历 x 中的字幕行。
    循环 $j = 1$ 至 k **执行** // 采样多个翻译候选。
      采样 $\pi_{\text{sft}}(t_i^j \mid x, t_1^{(\text{c})} \ldots, t_{i-1}^{(\text{c})})$.
    结束循环
    对候选集 $\{t_i^j \mid j = 1, \ldots, k\}$去重。
    （可选）将人工参考 t_i^0添加到候选集中，得到 $\mathcal{T}_i = \{t_i^j \mid j = 0, 1, \ldots\}$。
    通过 π_{e}评估 $\mathcal{T}_i$，得到分数序列 $\mathcal{E}_i$。
    选择一个更优的翻译 $t_i^{(\text{c})}$（从前 3 个中随机选择）作为下一轮采样循环的前缀。
  结束循环
结束循环
返回 $\mathcal{S}(x) \equiv \{(s_i, \mathcal{T}_i, \mathcal{E}_i) \mid i = 1, 2, \ldots, n\}$.

将 $\mathbf{1}(s_i)$ 设为 $= 0$；否则，将 $\mathbf{1}(s_i)$ 设为 $= 1$。重要性得分 $\delta(s_i)$ 取决于采样翻译的多样性。由于包含更丰富译文的行能为生动性增强提供更多可能，因此，对于去重后具有更多不同采样翻译的行，我们赋予更高的 $\delta(s_i)$：

$$\delta(s_i) = \frac{|\mathcal{T}_i|}{\sum_{j=1}^{n} |\mathcal{T}_j|}. \tag{3}$$

随后，我们采用 Bradley–Terry（DPO 风格）的偏好优化损失，实现翻译模型的细粒度偏好对齐：

$$\mathcal{L}_{\text{alpo}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, \mathcal{S}(x)) \sim \mathbb{D}_{\text{alpo}}} \left[\sum_{i=1}^{n} w(s_i) \cdot \log \sigma \left(\beta_i \log \frac{\pi_\theta(t_i^{(\text{c})} \mid p_i)}{\pi_{\text{ref}}(t_i^{(\text{c})} \mid p_i)} - \beta_i \log \frac{\pi_\theta(t_i^{(\text{r})} \mid p_i)}{\pi_{\text{ref}}(t_i^{(\text{r})} \mid p_i)} \right) \right]. \tag{4}$$

其中，π_θ 是策略模型，π_{ref} 是参考模型。选中译文 $t_i^{(\text{c})}$ 从 $\mathcal{E}_i$中得分最高的 3 个译文里随机抽取，而拒绝译文 $t_i^{(\text{r})}$ 则选取得分第三低的候选项（排除最低分项，以避免对比样本过于简单）。β_i是一个控制对奖励差异敏感度的超参数。由于 $t_i^{(\text{c})}$ 和 $t_i^{(\text{r})}$ 之间的奖励差距也反映了译文的多样性，我们按如下方式动态设置 β_i：

$$\beta_i = \frac{r(s_i, t_i^{(\text{c})}) - r(s_i, t_i^{(\text{r})})}{\max\{r(s_j, t_j^{(\text{c})}) - r(s_j, t_j^{(\text{r})}) \mid j \in [n]\}}, \tag{5}$$

其中，$r(s_i, t_i^{(\text{c})})$表示 $\mathcal{E}_i$中 $t_i^{(\text{c})}$ 的得分。此外，我们没有直接将$\mathcal{L}_{\text{alpo}}$中的前缀设为 $p_i = x, t_1^{(\text{c})}, \ldots, t_{i-1}^{(\text{c})}$，而是采用预定式前缀混合策略，以减轻推理过程中生成非选中译文时的曝光偏差。具体而言，以 λ的概率将 $t_i^{(\text{c})}$ 附加到前缀，以 $1 - \lambda$的概率将从 $\mathcal{T}_i$ 中采样的译文附加到前缀（在实验中，随着训练推进，λ 从 0.2 增至 0.6），即：

$$p_i = x, \hat{t}_1, \ldots, \hat{t}_{i-1}, \quad \hat{t}_j \leftarrow \text{Mix}(t_j^{(\text{c})}, t_j \sim \mathcal{T}, \lambda). \tag{6}$$

5 实验

5.1 实验设置

我们使用定制的 MuSC 数据集进行实验，该数据集包含来自在线视频平台优酷的多种方向的字幕语料库，包括 en⇒de、en⇒fr、en⇒zh、ko⇒zh、zh⇒en 和 zh⇒th。每种方向涵盖不同类型的 100–200 个节目，其中 10% 的节目作为测试集。MuSC 的详细信息见附录 A。

我们将模型与以下基线系统进行比较：

- **VideoDubber**（Wu 等人，2023）构建了一个长度受控翻译模型（这项工作与我们新提出的字幕翻译任务最为相关）。
- **NLLB-3.3B**（Costa-jussà 等人，2022）是 Meta AI 的翻译模型（>200 种语言）。
- **MADLAD-10B**（Kudugunta 等人，2024）是 Google 的翻译模型（>450 种语言）。
- **谷歌翻译**（Google LLC，2023）是 Google 的多语言翻译服务。
- **GPT-4o、通义千问-Max和DeepSeek-V3.1**（Liu 等人，2024）是当前最先进的聊天模型。
- **DeepSeek-R1** 和**GPT-5 Thinking** 是最先进的推理模型。我们使用附录 F.1 中的提示来驱动这些大语言模型进行字幕翻译。
- **Qwen2.5-14B** 作为 ALPO 的骨干网络。

实验设置的详细信息见附录 B。ALPO 的源代码和 MuSC 数据集可在 https://github.com/CcQunResearch/ALPO 获取。

5.2 翻译质量评估

与法律文本和技术文档翻译对准确性的严格要求不同，字幕翻译更注重翻译的本地化表达（Pérez-González，2019年；Baǰci´c & Golenko，2024）。具体而言，字幕翻译不仅要求传达原意，还应体现原字幕的情感和语气，力求使译文更加流畅、生动且富有表现力。鉴于我们已在第3.1节验证了大语言模型作为评估器的可靠性，因此基于大语言模型评审评分，为字幕翻译构建了多维质量评估系统。我们主要从三个维度评估字幕翻译：（1）准确性：译文是否准确传达了字幕原意；（2）自然度：译文表达是否自然流畅，是否符合目标语言的语法结构和词汇习惯；（3）生动性：译文是否富有表现力，能否成功传达原字幕的情感和氛围。我们采用 DeepSeek-V3.1、Claude Sonnet 4 和 GPT-5 Instant 作为评估模型，对测试集中的字幕片段从这三个维度进行评分，每个维度的分数范围为 0 至 100。随后，我们在表 3 中展示所有模型的平均分。

表 3：多维度质量评估。第 1 和第 2 名的最佳结果分别用**蓝色**和**橙色**表示。ST 指监督训练。指学习。

模型	训练	en⇒de			en⇒fr			en⇒zh		
		准确性	自然度	生动性	准确性	自然度	生动性	准确性	自然度	生动性
黄金参考	人工	84.8	83.8	**73.1**	83.5	85.0	74.8	83.6	82.6	**71.5**
VideoDubber	ST	41.5	35.5	41.0	48.2	47.3	48.5	46.9	51.9	49.7
NLLB-3.3B	ST	75.8	70.1	59.2	76.9	74.6	61.8	61.4	54.0	43.7
MADLAD-10B	ST	73.2	65.6	54.5	73.6	67.8	57.4	59.7	55.5	46.3
谷歌翻译	ST	89.9	80.6	62.6	91.9	84.8	64.3	84.2	79.7	54.4
GPT-4o	ICL（C）	94.1	86.7	66.9	93.2	88.8	69.5	89.3	82.3	59.8
通义千问-Max	ICL（C）	**95.5**	**89.2**	68.9	**94.1**	89.9	71.6	**91.9**	84.4	61.3
DeepSeek-V3.1	ICL（C）	94.8	**89.2**	67.4	**94.9**	**90.7**	72.3	91.2	85.3	63.5
DeepSeek-R1	ICL（R）	92.8	87.4	70.0	93.6	90.0	73.4	90.5	**85.7**	70.8
GPT-5	ICL（R）	93.6	88.6	72.7	92.2	**90.4**	**75.8**	**92.4**	**87.0**	71.1
Qwen2.5-14B	SFT	87.5	83.6	64.4	86.2	85.4	67.7	86.4	82.0	59.1
Qwen2.5-14B	**ALPO**	**95.4**	88.4	**74.8**	94.1	89.2	**78.8**	90.6	84.3	**76.6**

模型	训练	韩语⇒中文			中文⇒英语			中文⇒泰语		
		准确性	自然度	生动性	准确性	自然度	生动性	准确性	自然度	生动性
黄金参考	人工	78.0	77.8	**65.8**	83.0	80.3	73.3	76.6	75.1	66.3
VideoDubber	ST	39.6	45.2	48.2	53.6	54.8	50.1	34.1	34.9	41.5
NLLB-3.3B	ST	33.1	26.1	25.4	29.1	21.7	20.8	42.6	33.9	40.5
MADLAD-10B	ST	44.9	42.9	46.7	45.1	38.9	37.6	47.9	50.8	51.0
谷歌翻译	ST	54.9	52.8	52.0	79.8	66.3	50.2	55.2	56.2	54.5
GPT-4o	ICL（C）	80.0	79.9	58.1	88.5	83.0	64.6	88.0	84.4	67.9
通义千问-Max	ICL（C）	83.7	82.5	61.8	**90.0**	85.0	66.8	**91.3**	**85.8**	69.1
DeepSeek-V3.1	ICL（C）	83.1	82.2	57.2	**89.5**	84.1	63.0	89.9	84.6	67.1
DeepSeek-R1	ICL（R）	79.8	81.6	65.6	88.5	85.6	73.5	87.6	84.0	71.0
GPT-5	ICL（R）	**84.5**	**82.6**	65.0	89.1	**86.1**	**75.2**	88.7	83.9	**73.0**
Qwen2.5-14B	SFT	80.9	76.1	53.9	85.2	80.1	54.8	87.3	82.6	66.0
Qwen2.5-14B	**ALPO**	**84.3**	**83.3**	**70.5**	88.3	**86.8**	**81.7**	**91.9**	**84.7**	**74.2**

结果表明，ALPO 和 GPT-4o 等先进的大语言模型显著优于 MADLAD 等传统翻译模型，并且在准确性和自然度方面超过了人工翻译。然而，人工翻译在生动性方面表现更出色，这或许说明人工译文通常会融入更多意译，并结合视频内容进行翻译。与聊天模型相比，推理模型的生动性评分更高，但其准确性和自然度略低。

在某些方面如此。这与第 3.3 节的结论一致。与 SFT模型相比，采用 ALPO 训练的模型不仅在生动性方面显著提升，在准确性和自然度方面也有所改善。ALPO 在所有方向上的生动性评分均为最高，并且在其他维度上偶尔也超过了其他先进的大语言模型，尤其是在 ko⇒zh 和 zh⇒th 等相对低资源的语言方向上表现突出。质量评估的完整结果以及大语言模型评估器的提示设计，请参阅附录 C.1 和附录 F.3。

5.3 翻译质量的人工评估

我们在表 4 中展示了 ALPO 在 en⇒zh 和 zh⇒th 方向上的翻译质量人工评估结果。由于评估者的主观偏好各不相同，我们不采用评分系统，而是通过对不同译文进行成对比较来评估胜率指标。我们从准确性、自然度和生动性等维度评估 ALPO 模型和四个基线系统，并进行全面评估。人工评估结果与表 3 所示结果一致，这支持了我们基于 LLM-as-a-Judge 的多维评估系统的可靠性。具体实验设置见附录 B.3。

表 4：人工翻译质量评估，报告胜率（胜：平：负）。获胜和落败的对比项分别以蓝色和橙色标示。

挑战方	竞争方	en⇒zh 准确性	en⇒zh 自然度	en⇒zh 生动性	en⇒zh 全面性	zh⇒th 准确性	zh⇒th 自然度	zh⇒th 生动性	zh⇒th 全面性
ALPO (Qwen2.5-14B)	黄金参考	29:49:22	28:50:22	32:42:26	31:46:23	28:48:24	26:52:22	32:39:29	29:49:22
	SFT模型 π_{sft}	26:50:24	31:48:21	38:41:21	37:43:20	26:55:19	27:51:22	39:42:19	30:50:20
	GPT-4o	22:54:24	20:57:23	29:51:20	26:54:23	23:57:20	23:51:26	36:37:27	30:45:25
	DeepSeek-R1	22:55:23	19:57:24	22:58:20	20:59:21	23:55:22	18:62:20	28:50:22	27:47:24

5.4 消融研究

5.4.1 自适应策略

我们在表 5 中验证了 ALPO 中几种自适应策略对模型性能的影响。结果表明，门控函数 $\mathbf{1}(s_i)$ 的影响最大，这说明门控机制能有效减少低多样性无关语句带来的噪声（例如“早上好”这类简单语句）。此外，移除重要性得分 $\delta(s_i)$、动态 β 和计划式前缀混合，也会导致性能出现不同程度的下降，凸显了训练过程中进行细粒度控制的必要性。

表 5：自适应策略的影响。

	en⇒zh 准确性	en⇒zh 自然度	en⇒zh 生动性	zh⇒th 准确性	zh⇒th 自然度	zh⇒th 生动性
SFT	*86.5*	*82.1*	*59.2*	*87.2*	*82.4*	*66.0*
ALPO	*90.6*	*84.2*	*76.6*	*91.9*	*84.7*	*74.1*
无 $w\ (s_i)$	88.1	83.2	67.4(↓9.2)	89.2	83.4	70.3(↓3.8)
无 $\mathbf{1}(s_i)$	89.1	83.4	70.2(↓6.4)	89.9	83.9	71.2(↓2.9)
无 $\delta(s_i)$	89.4	83.5	72.4(↓4.2)	90.2	83.1	71.9(↓2.2)
固定 β_i ($\beta = 0.5$)	90.0	83.4	74.7(↓1.9)	89.8	83.4	72.1(↓2.0)
不使用前缀混合	89.5	83.7	73.4(↓3.2)	88.8	83.7	71.6(↓2.5)

5.4.2 骨干模型

我们在表6中比较了不同骨干模型的性能。结果表明，与SFT模型（Qwen2.5-14B）相比，ALPO 能使所有骨干模型的性能显著提升。值得注意的是，参数更多、规模更大的 Qwen2.5-14B 模型表现更优且稳定。LLaMA-3.1-8B 在中文相关任务上的表现不佳，因此在 en⇒zh 和 zh⇒en 两个翻译方向上的结果均逊于另外两个模型。尽管如此，与 SFT基线相比，它仍取得了明显的提升。

Tab表 6：骨干模型的影响。第 1 名和第 2 名的结果分别标记为 **蓝色** 和 **橙**.

方法	骨干网络	en⇒de 准确性	en⇒de 自然度	en⇒de 生动性	en⇒zh 准确性	en⇒zh 自然度	en⇒zh 生动性	zh⇒en 准确性	zh⇒en 自然度	zh⇒en 生动性
SFT	-	*87.7*	*83.4*	*64.4*	*86.5*	*82.1*	*59.2*	*85.2*	*80.1*	*54.9*
ALPO	LLaMA-3.1-8B（Dubey 等，2024）	**94.7**	**87.2**	**73.9**	88.0	83.2	72.3	87.2	85.3	77.6
	GLM4-9B（GLM 等，2024）	93.2	86.1	73.2	**88.2**	**84.4**	**74.4**	**89.2**	**85.7**	**78.2**
	Qwen2.5-14B（Yang 等，2024a）	**95.2**	**88.3**	**74.8**	**90.6**	**84.2**	**76.6**	**88.3**	**86.8**	**81.6**

5.5 其他实验

我们在附录中进行了以下扩展实验：1）在附录 C.1 中，我们展示了质量评估的完整结果，并进行了进一步分析。2）在附录 C.2 中，我们展示了 ALPO 译文示例。3）在附录 C.3 中，我们针对采样、模型规模等其他因素进行了消融研究。4）在附录 C.4 中，我们进行了性能评估，对比 DPO 和近端策略优化（PPO）。5）在附录 C.5 中，我们验证了 ALPO 在其他应用任务中的有效性。

6 结论

在本研究中，我们考察了字幕翻译以及其他领域中的直译与意译，并验证了大语言模型作为翻译质量评估器和奖励模型的可靠性。在此基础上，我们提出了用于训练富有表现力且生动的字幕翻译的自适应局部偏好优化（ALPO）方法。在既定的翻译质量评估框架下开展的实验验证了 ALPO 的有效性。

伦理声明

本研究仅使用从在线视频平台优酷获取的字幕数据。本研究所使用的全部数据均通过适当授权获得，并遵守优酷制定的条款和条件。数据仅用于学术目的，并以确保符合相关道德和法律标准的方式进行处理。

本研究未收集或处理任何个人信息或敏感信息。所有分析均基于优酷提供的可访问内容进行，我们未尝试推断或披露材料中涉及的任何个人或实体的隐私信息或身份信息。

我们确认，本研究遵循负责任且符合伦理的研究原则，并遵守所有适用的机构和监管指南。

可复现性声明

为便于复现并推动进一步研究，我们公开了本研究使用的所有代码、训练脚本和 MuSC 数据集。详细的实验设置、模型配置和评估协议见正文及附录。源代码和数据集可通过 https://github.com/CcQunResearch/ALPO 公开获取。

致谢

作者谨向匿名审稿人表示感谢，感谢他们提出的宝贵意见和建议，这些意见和建议提升了本文的质量。

参考文献

Atef Odeh AbuSa'aleek. 机器翻译在翻译伊斯兰教文本方面的充分性与可接受性。《国际英语语言学杂志》，6：185–193，2016年。

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。《GPT-4 技术报告》。arXiv 预印本 arXiv:2303.08774，2023。

Zhiyu An、Xianzhong Ding、Yen-Chun Fu、Cheng-Chung Chu、Yan Li 和 Wan Du。《Golden-retriever：面向工业知识库的高保真智能体检索增强生成》。arXiv 预印本 arXiv:2408.00798，2024。

Marcin Andrychowicz、Anton Raichuk、Piotr Sta´nczyk、Manu Orsini、Sertan Girgin、Raphaël Marinier、Leonard Hussenot、Matthieu Geist、Olivier Pietquin、Marcin Michalski 等人。同策略深度演员-评论家方法的关键因素是什么？一项大规模研究。载于《国际学习表征会议》，2021 年。

Anthropic。Anthropic 透明度中心：模型报告，2025 年。网址：https://www.anthropic.com/transparency/model-report。

Mohammad Gheshlaghi Azar、Zhaohan Daniel Guo、Bilal Piot、Remi Munos、Mark Rowland、Michal Valko 和 Daniele Calandriello。从人类偏好中学习的一般理论范式。载于《国际人工智能与统计会议》，第 4447–4455 页。PMLR，2024 年。

Yushi Bai、Jiahao Ying、Yixin Cao、Xin Lv、Yuze He、Xiaozhi Wang、Jifan Yu、Kaisheng Zeng、Yijia Xiao、Haozhe Lyu 等。《使用语言模型作为审查者对基础模型进行基准测试》。神经信息处理系统进展，36：78142–78167，2023。

Martina Bajˇci´c 和 Dejana Golenko。《在法律翻译中应用大语言模型：现状》。语言与法律国际期刊（JLL），13：171–196，2024。

Mona Baker 和 Gabriela Saldanha。《Routledge翻译研究百科全书》。Routledge，2019年。

Susan Bassnett。《翻译研究》。Routledge，2013年。

Emily M Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。《论随机鹦鹉的危害：语言模型会不会过于庞大？》。载于《2021年ACM公平、问责与透明会议论文集》，第610–623页，2021年。

Rishi Bommasani、Drew A Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S Bernstein、Jeannette Bohg、Antoine Bosselut、Emma Brunskill 等。《基础模型的机遇与风险》。arXiv 预印本 arXiv:2108.07258，2021。

Daoyuan Chen、Yilun Huang、Zhijian Ma、Hesen Chen、Xuchen Pan、Ce Ge、Dawei Gao、Yuexiang Xie、Zhaoyang Liu、Jinyang Gao 等。Data-juicer：面向大语言模型的一站式数据处理系统。载于《2024年数据管理国际会议伴会论文集》，第120–134页，2024a。

Dongping Chen、Ruoxi Chen、Shilin Zhang、Yaochen Wang、Yinuo Liu、Huichi Zhou、Qihui Zhang、Yao Wan、Pan Zhou 和 Lichao Sun。多模态大语言模型评审：使用视觉语言基准评估多模态大语言模型评审。载于《第四十一届国际机器学习大会》，2024b。

Wei-Lin Chiang、Zhuohan Li、Zi Lin、Ying Sheng、Zhanghao Wu、Hao Zhang、Lianmin Zheng、Siyuan Zhuang、Yonghao Zhuang、Joseph E Gonzalez 等。《Vicuna：一款以 90%* ChatGPT 质量惊艳 GPT-4 的开源聊天机器人》。参见 https://vicuna. lmsys. org（访问于 2023年4月14日），2(3):6，2023年。Paul F Christiano、Jan Leike、Tom Brown、Miljan Martic、Shane Legg 和 Dario Amodei。《从人类偏好中进行深度强化学习》。神经信息处理系统进展，30，2017年。Christos Christodouloupoulos 和 Mark Steedman。《大规模平行语料库：100 种语言的《圣经》》。《语言资源与评估》，49:375–395，2015年。Marta R Costa-jussà、James Cross、Onur Çelebi、Maha Elbayad、Kenneth Heafield、Kevin Heffernan、Elahe Kalbassi、Janice Lam、Daniel Licht、Jean Maillard 等。《不让任何语言掉队：扩展以人为中心的机器翻译》。arXiv 预印本 arXiv:2207.04672，2022年。Hanze Dong、Wei Xiong、Deepanshu Goyal、Yihan Zhang、Winnie Chow、Rui Pan、Shizhe Diao、Jipeng Zhang、Kashun Shum 和 Tong Zhang。《RAFT：用于生成式基础模型对齐的奖励排序微调》。《机器学习研究汇刊》，2023年，2023年。

Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Amy Yang、Angela Fan 等。《Llama 3 模型群》。arXiv 预印本 arXiv:2407.21783，2024年。Logan Engstrom、Andrew Ilyas、Shibani Santurkar、Dimitris Tsipras、Firdaus Janoos、Larry Rudolph 和 Aleksander Madry。《深度策略梯度中的实现至关重要：关于近端策略优化和信赖域策略优化的案例研究》。载于《国际学习表征会议》，2020年。Maxim Enis 和 Mark Hopkins。《从大语言模型到神经机器翻译：借助 Claude 推进低资源机器翻译》。arXiv 预印本 arXiv:2404.13813，2024年。Kawin Ethayarajh、Winnie Xu、Niklas Muennighoff、Dan Jurafsky 和 Douwe Kiela。《KTO：将模型对齐视为前景理论优化》。arXiv 预印本 arXiv:2402.01306，2024年。Marcello Federico、Robert Enyedi、Roberto Barra-Chicote、Ritwik Giri、Umut Isik、Arvindh Krishnaswamy 和 Hassan Sawaf。《从语音到语音翻译到自动配音》。载于第 17 届《国际口语翻译会议》论文集，第 257–264 页，2020年。Zhaopeng Feng、Ruizhe Chen、Yan Zhang、Zijie Meng 和 Zuozhu Liu。《Ladder：一个模型无关框架，助力基于大语言模型的机器翻译更上一层楼》。载于《2024年自然语言处理实证方法会议》论文集，第 15377–15393 页，2024年。GLM 团队、Aohan Zeng、Bin Xu、Bowen Wang、Chenhui Zhang、Da Yin、Dan Zhang、Diego Rojas、Guanyu Feng、Hanlin Zhao 等。《ChatGLM：从 GLM-130B 到 GLM-4 All Tools 的大语言模型系列》。arXiv 预印本 arXiv:2406.12793，2024年。Google LLC。《谷歌翻译》，2023年。URL https://translate.google.com。访问日期：2025-05-15。Jiawei Gu、Xuhui Jiang、Zhichao Shi、Hexiang Tan、Xuehao Zhai、Chengjin Xu、Wei Li、Yinghan Shen、Shengjie Ma、Honghao Liu 等。《LLM-as-a-Judge 综述》。arXiv 预印本 arXiv:2411.15594，2024年。Nuno M Guerreiro、Ricardo Rei、Daan van Stigt、Luisa Coheur、Pierre Colombo 和 André FT Martins。《XCOMET：通过细粒度错误检测实现透明的机器翻译评估》。《计算语言学协会会刊》，12:979–995，2024年。Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。《DeepSeek-R1：通过强化学习激励大语言模型的推理能力》。arXiv 预印本 arXiv:2501.12948，2025年。Kaiyu Huang、Fengran Mo、Xinyu Zhang、Hongliang Li、You Li、Yuanchi Zhang、Weijian Yi、Yulong Mao、Jinchen Liu、Yuzhuang Xu 等。《多语言大语言模型综述：最新进展与新前沿》。arXiv 预印本 arXiv:2405.10936，2024年。Yunjie Ji、Yong Deng、Yan Gong、Yiping Peng、Qiang Niu、Baochang Ma 和 Xiangang Li。《BELLE：打造人人可用的大语言模型引擎》，2023年。伊曼努尔·康德。《纯粹理性批判》。1781年。《现代古典哲学家》，马萨诸塞州剑桥：霍顿·米夫林，第 370–456 页，1908年。伊曼努尔·康德。《判断力批判》。Hackett 出版社，1987年。Pei Ke、Bosi Wen、Andrew Feng、Xiao Liu、Xuanyu Lei、Jiale Cheng、Shengyuan Wang、Aohan Zeng、Yuxiao Dong、Hongning Wang 等。《CritiqueLLM：构建用于评估大语言模型生成结果的信息丰富的批评生成模型》。载于第 62 届《计算语言学协会年会》（第 1 卷：长文）论文集，第 13034–13054 页，2024年。Muhammad Uzair Khattak、Hanoona Rasheed、Muhammad Maaz、Salman Khan 和 Fahad Shahbaz Khan。《MAPLE：多模态提示学习》。载于《IEEE/CVF 计算机视觉与模式识别会议》论文集，第 19113–19122 页，2023年。

Aobo Kong、Wentao Ma、Shiwan Zhao、Yongbin Li、Yuchuan Wu、Ke Wang、Xiaoqian Liu、Qicheng Li、Yong Qin 和 Fei Huang。SDPO：面向社交智能体的片段级直接偏好优化。arXiv 预印本 arXiv:2501.01821，2025年。Sneha Kudugunta、Isaac Caswell、Biao Zhang、Xavier Garcia、Derrick Xin、Aditya Kusupati、Romi Stella、Ankur Bapna 和 Orhan Firat。MADLAD-400：一个多语言、文档级的大型审计数据集。《神经信息处理系统进展》，36，2024年。Junlong Li、Shichao Sun、Weizhe Yuan、Run-Ze Fan、Pengfei Liu 等。用于评估对齐的生成式评判器。载于《第十二届国际学习表征会议》，2024a。Lijun Li、Bowen Dong、Ruohui Wang、Xuhao Hu、Wangmeng Zuo、Dahua Lin、Yu Qiao 和 Jing Shao。SALAD-bench：面向大语言模型的分层全面安全基准。载于《计算语言学协会：ACL 2024》论文集，3923–3954页，2024b。Chen Ling、Xujiang Zhao、Jiaying Lu、Chengyuan Deng、Can Zheng、Junxiang Wang、Tanmoy Chowdhury、Yun Li、Hejie Cui、Xuchao Zhang 等。领域专业化是推动大语言模型产生颠覆性影响的关键：一项全面综述。arXiv 预印本 arXiv:2305.18703，2023。Pierre Lison 和 Jörg Tiedemann。OpenSubtitles2016：从电影和电视字幕中提取大型平行语料库。载于第十届《语言资源与评估》国际会议（LREC'16）论文集，923–929页，2016年。Aiwei Liu、Haoping Bai、Zhiyun Lu、Yanchao Sun、Xiang Kong、Xiaoming Simon Wang、Jiulong Shan、Albin Madappally Jose、Xiaojiang Liu、Lijie Wen 等。TIS-DPO：使用估计权重进行直接偏好优化的词元级重要性采样。载于《第十三届国际学习表征会议》，2025年。Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan 等。DeepSeek-V3 技术报告。arXiv 预印本 arXiv:2412.19437，2024年。Keming Lu、Bowen Yu、Chang Zhou 和 Jingren Zhou。大语言模型是所有角色的叠加：通过自对齐实现任意角色扮演。arXiv 预印本 arXiv:2401.12474，2024年。Haipeng Luo、Qingfeng Sun、Can Xu、Pu Zhao、Jianguang Lou、Chongyang Tao、Xiubo Geng、Qingwei Lin、Shifeng Chen 和 Dongmei Zhang。WizardMath：通过强化进化指令增强大语言模型的数学推理能力。arXiv 预印本 arXiv:2308.09583，2023年。Trung Quoc Luong、Xinbo Zhang、Zhanming Jie、Peng Sun、Xiaoran Jin 和 Hang Li。ReFT：通过强化微调进行推理。arXiv 预印本 arXiv:2401.08967，3，2024年。James MacGlashan、Mark K Ho、Robert Loftin、Bei Peng、Guan Wang、David L Roberts、Matthew E Taylor 和 Michael L Littman。从策略依赖的人类反馈中进行交互式学习。载于《国际机器学习会议》，2285–2294页。PMLR，2017年。Yu Meng、Mengzhou Xia 和 Danqi Chen。SimPO：一种无需参考模型的简单偏好优化方法。arXiv 预印本 arXiv:2405.14734，2024年。Shivam Mhaskar、Nirmesh Shah、Mohammadi Zaki、Ashishkumar Gudmalwar、Pankaj Wasnik 和 Rajiv Shah。基于音素计数比奖励的强化学习等长神经机器翻译。载于《计算语言学协会：NAACL 2024》论文集，3966–3976页，2024年。Jeremy Munday、Sara Ramos Pinto 和 Jacob Blakesley。《翻译研究导论：理论与应用》。Routledge，2022年。

OpenAI。Gpt-5，2025年。网址：https://openai.com/gpt-5。

Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。《利用人类反馈训练遵循指令的语言模型》。神经信息处理系统进展，35：27730–27744，2022年。Jianhui Pang、Fanghua Ye、Derek Fai Wong、Dian Yu、Shuming Shi、Zhaopeng Tu 和 Longyue Wang。《向经典致敬：重审大语言模型时代机器翻译面临的挑战》。计算语言学协会会刊，13：73–95，2025年。Luis Pérez-González。视听翻译。载于《Routledge翻译研究百科全书》，第30–34页。Routledge，2019年。Pranav Putta、Edmund Mills、Naman Garg、Sumeet Motwani、Chelsea Finn、Divyansh Garg 和 Rafael Rafailov。《Agent Q：自主人工智能智能体的高级推理与学习》。arXiv预印本 arXiv:2408.07199，2024年。Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon 和 Chelsea Finn。《直接偏好优化：你的语言模型其实暗中充当奖励模型》。神经信息处理系统进展，36，2024年。Eman Rashed Alkatheery。《谷歌翻译在法律文本中的错误：机器翻译质量评估》。AWEJ翻译与文学研究，7(1)，2023年。Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Y Wu 等。《Deepseekmath：推动开放语言模型数学推理能力的极限》。arXiv预印本 arXiv:2402.03300，2024年。Wentao Shi、Mengqi Yuan、Junkang Wu、Qifan Wang 和 Fuli Feng。《面向语言智能体的直接多轮偏好优化》。载于《2024年自然语言处理实证方法会议论文集》，第2312–2324页，2024年。Noah Shinn、Federico Cassano、Ashwin Gopinath、Karthik Narasimhan 和 Shunyu Yao。《Reflexion：采用语言强化学习的语言智能体》。神经信息处理系统进展，36：8634–8652，2023年。Yifan Song、Da Yin、Xiang Yue、Jie Huang、Sujian Li 和 Bill Yuchen Lin。《试错：基于探索的大语言模型智能体轨迹优化》。载于《计算语言学协会第62届年会论文集（第一卷：长篇论文）》，第7584–7600页，2024年。Ralf Steinberger、Bruno Pouliquen、Anna Widiger、Camelia Ignat、Tomaz Erjavec、Dan Tufis 和 Dániel Varga。《JRC-Acquis：包含 20+种语言的多语言对齐平行语料库》。arXiv预印本 cs/0609058，2006年。Ralf Steinberger、Andreas Eisele、Szymon Klocek、Spyridon Pilos 和 Patrick Schlüter。《DGT-TM：一种免费提供的22种语言翻译记忆库》。arXiv预印本 arXiv:1309.5226，2013年。Ralf Steinberger、Mohamed Ebrahim、Alexandros Poulis、Manuel Carrasco-Benitez、Patrick Schlüter、Marek Przybyszewski 和 Signe Gilbro。《欧盟高度多语言平行语料库概述》。《语言资源与评估》，48：679–707，2014年。Nisan Stiennon、Long Ouyang、Jeffrey Wu、Daniel Ziegler、Ryan Lowe、Chelsea Voss、Alec Radford、Dario Amodei 和 Paul F Christiano。《利用人类反馈学习摘要生成》。神经信息处理系统进展，33：3008–3021，2020。Zhiqing Sun、Sheng Shen、Shengcao Cao、Haotian Liu、Chunyuan Li、Yikang Shen、Chuang Gan、Liangyan Gui、Yu-Xiong Wang、Yiming Yang 等。《通过事实增强的基于人类反馈的强化学习对齐大型多模态模型》。载于《计算语言学协会 ACL 2024 年会论文集》，第13088–13110页，2024年。Richard S Sutton。《强化学习：导论》。A Bradford Book，2018年。Rohan Taori、Ishaan Gulrajani、Tianyi Zhang、Yann Dubois、Xuechen Li、Carlos Guestrin、Percy Liang 和 Tatsunori B Hashimoto。《Stanford Alpaca：一个遵循指令的 Llama 模型》，2023年。

Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kulshreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du等。LaMDA：面向对话应用的语言模型。arXiv预印本 arXiv:2201.08239，2022。Jörg Tiedemann。OPUS中的平行数据、工具与接口。载于Nicoletta Calzolari（会议主席）、Khalid Choukri、Thierry Declerck、Mehmet Ugur Dogan、Bente Maegaard、Joseph Mariani、Jan Odijk与Stelios Piperidis（编），《第八届语言资源与评估国际会议论文集》（LREC'12），土耳其伊斯坦布尔，2012年5月。欧洲语言资源协会（ELRA）。ISBN 978-2-9517408-7-7。Ruiyi Wang、Haofei Yu、Wenxin Zhang、Zhengyang Qi、Maarten Sap、Graham Neubig、Yonatan Bisk与Hao Zhu。SOTOPIA-π：社会智能语言智能体的交互式学习。arXiv预印本 arXiv:2403.08715，2024a。Tianlu Wang、Ilia Kulikov、Olga Golovneva、Ping Yu、Weizhe Yuan、Jane Dwivedi-Yu、Richard Yuanzhe Pang、Maryam Fazel-Zarandi、Jason Weston与Xian Li。自教评估器。arXiv预印本 arXiv:2408.02666，2024b。Yidong Wang、Zhuohao Yu、Wenjin Yao、Zhengran Zeng、Linyi Yang、Cunxiang Wang、Hao Chen、Chaoya Jiang、Rui Xie、Jindong Wang等。PandaLM：面向大语言模型指令微调优化的自动评估基准。载于第十二届国际学习表征会议，2024c。Zilong Wang、Zifeng Wang、Long Le、Huaixiu Steven Zheng、Swaroop Mishra、Vincent Perot、Yuwei Zhang、Anush Mattapalli、Ankur Taly、Jingbo Shang等。推测式RAG：通过起草增强检索增强生成。arXiv预印本 arXiv:2407.08223，2024d。Yihan Wu、Junliang Guo、Xu Tan、Chen Zhang、Bohan Li、Ruihua Song、Lei He、Sheng Zhao、Arul Menezes与Jiang Bian。VideoDubber：面向视频配音的语音感知长度控制机器翻译。载于AAAI人工智能会议论文集，第37卷，第13772–13779页，2023。Tianyi Xiong、Xiyao Wang、Dong Guo、Qinghao Ye、Haoqi Fan、Quanquan Gu、Heng Huang与Chunyuan Li。LLaVA-Critic：学习评估多模态模型。arXiv预印本 arXiv:2410.02712，2024。Haoran Xu、Young Jin Kim、Amr Sharaf与Hany Hassan Awadalla。机器翻译的范式转变：提升大语言模型的翻译性能。载于第十二届国际学习表征会议，2024a。Haoran Xu、Kenton Murray、Philipp Koehn、Hieu Hoang、Akiko Eriguchi与Huda Khayrallah。X-ALMA：用于大规模高质量翻译的即插即用模块与自适应拒绝。arXiv预印本 arXiv:2410.03115，2024b。Haoran Xu、Amr Sharaf、Yunmo Chen、Weiting Tan、Lingfeng Shen、Benjamin Van Durme、Kenton Murray与Young Jin Kim。对比偏好优化：拓展大语言模型在机器翻译中的性能边界。载于第41届国际机器学习会议论文集，第55204–55224页，2024c。An Yang、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoran Wei等。Qwen2.5技术报告。arXiv预印本 arXiv:2412.15115，2024a。An Yang、Beichen Zhang、Binyuan Hui、Bofei Gao、Bowen Yu、Chengpeng Li、Dayiheng Liu、Jianhong Tu、Jingren Zhou、Junyang Lin等。Qwen2.5-Math技术报告：通过自我改进迈向数学专家模型。arXiv预印本 arXiv:2409.12122，2024b。An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv等。Qwen3技术报告。arXiv预印本 arXiv:2505.09388，2025。Yongcheng Zeng、Guoqing Liu、Weiyu Ma、Ning Yang、Haifeng Zhang与Jun Wang。词元级直接偏好优化。载于第41届国际机器学习会议论文集，第58348–58365页，2024。

Beichen Zhang、Kun Zhou、Xilin Wei、Xin Zhao、Jing Sha、Shijin Wang与Ji-Rong Wen。评估并改进工具增强的计算密集型数学推理。《神经信息处理系统进展》，36：23570–23589，2023。

Ran Zhang、Wei Zhao 与 Steffen Eger。大语言模型在文学翻译方面究竟表现如何？基于人类与大语言模型的文学翻译评估。arXiv 预印本 arXiv:2410.18697，2024。

Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric Xing 等。使用 MT-Bench 与 Chatbot Arena 评判 LLM-as-a-Judge。神经信息处理系统进展，36:46595–46623，2023a。

Rui Zheng、Shihan Dou、Songyang Gao、Yuan Hua、Wei Shen、Binghai Wang、Yan Liu、Senjie Jin、Qin Liu、Yuhao Zhou 等。大语言模型中基于人类反馈的强化学习的秘密（第一部分）：近端策略优化。arXiv 预印本 arXiv:2307.04964，2023b。

Xuhui Zhou、Hao Zhu、Leena Mathur、Ruohong Zhang、Haofei Yu、Zhengyang Qi、Louis-Philippe Morency、Yonatan Bisk、Daniel Fried、Graham Neubig 等。Sotopia：语言智能体社交智能的交互式评估。载于《第十三届国际学习表征会议》，2024。

Lianghui Zhu、Xinggang Wang 和 Xinlong Wang。JudgeLM：针对大语言模型进行微调的 arXiv 可扩展评审器。载于《第十三届国际学习表征会议》，2025。

Daniel M Ziegler、Nisan Stiennon、Jeffrey Wu、Tom B Brown、Alec Radford、Dario Amodei、Paul Christiano 和 Geoffrey Irving。基于人类偏好微调语言模型。arXiv 预印本 arXiv:1909.08593，2019年。

附录目录

附录章节	目录
附录 A	数据来源与数据集构建
附录 B	实验设置的补充细节
附录 C	ALPO 扩展评估实验
附录 D	学说：自适应局部偏好优化
附录 E	关于 ALPO 的进一步讨论
附录F	ALPO 中的提示与指令

A 数据集构建

本节将介绍模型训练所用数据集的相关详情。

A.1 数据来源

本研究使用在线视频平台优酷的 MuSC 数据集进行实验，该数据集包含平台节目中的源语言字幕和多语种译制字幕。表7展示了 MuSC 数据集中的英语和中文字幕样例（采用 ass 文件格式），其中“Start”和“End”列表示字幕在节目中的起始和结束时间。MuSC 涵盖的翻译方向包括 en⇒de、en⇒fr、en⇒zh、ko⇒zh、zh⇒en 和 zh⇒th。这些翻译方向涵盖跨语系和跨语支场景。字幕翻译是一项数据丰富型任务，因此领先的在线视频平台能够轻松收集大量多语言字幕语料库。MuSC 数据集在每个翻译方向均包含 100 多个节目，涵盖多年来的电影、电视剧、纪录片和动画等类型。MuSC 数据集的统计数据见表8。

与法律文本和医学等领域要求高度准确性的严肃文本翻译不同（AbuSa’aleek, 2016年；Rashed Alkatheery, 2023年；Bajˇci´c & Golenko, 2024），字幕

视觉媒体节目中的文本与视频和音频模态信息紧密相关，因此翻译时对准确性有所损失也有一定容忍度。然而，翻译字幕必须准确传达原文的情感、语气、氛围和文化上下文。因此，人工翻译并非只是直接翻译原字幕，还包括润色、改写和改编，以满足目标受众的本地化需求（Bassnett, 2013年；Pérez-González, 2019年）。因此，人工译者生成的黄金参考译文是经过润色的原文版本，这意味着用于训练的平行语料库并非真正“平行”（第 3.2 节将对此结论进行讨论），原文与译文之间存在信息不对称（示例见表7）。因此，在构建训练集时，每个方向的节目都是母语节目。

表7：多语言字幕示例。

开始	End	Text
0:17:33.25	0:17:34.50	拜托， 让我说！
0:17:39.54	0:17:42.83	作为下议院议员， 我的声音在这里没什么分量。
0:17:42.91	0:17:46.45	但作为一位母亲， 我的语态至关重要。
0:17:47.50	0:17:50.25	我的儿子神志不清。
0:17:51.04	0:17:53.95	他这一生， 都在追逐一个不可能实现的梦想。
0:17:54.91	0:17:59.00	他所做的事 ， uh， 愚蠢又不明智。
0:17:59.83	0:18:04.20	但他心地善良。 拜托， 让他回家。
0:18:04.29	0:18:08.66	这样的罪行不容忽视。 男孩必须受到惩罚。
0:18:08.75	0:18:12.16	违背 Ethos 者应被驱逐 ，
0:18:12.25	0:18:16.25	但我能理解一个年轻人改变世界的梦想。
0:18:16.83	0:18:20.58	或许在这件事上， 较轻的刑罚或许就足够了。
0:17:42.91	0:17:46.50	我以母亲的身份，恳请在座的各位议员听一听
0:17:47.00	0:17:50.25	我这儿子确实犯下了不可饶恕的错误
0:17:51.04	0:17:53.95	他这辈子都在追逐一个不可能实现的梦想
0:17:54.91	0:17:59.33	他的所作所为既愚蠢又不明智
0:17:59.83	0:18:04.25	但他并无恶意，求求你们，放他回家吧。
0:18:04.33	0:18:06.70	如此严重的罪行，不能轻易放过。
0:18:06.79	0:18:08.66	这小子必须受到惩罚。
0:18:08.75	0:18:12.20	违反社会共识的人确实应该遭到驱逐。
0:18:12.29	0:18:16.25	但我也能体会一个年轻人想要改变世界的雄心
0:18:16.33	0:18:20.58	就这个案件来说，还是酌情轻判为好

表8：数据集的统计数据。FM、TV、DO 和 AN 分别是电影、电视剧、纪录片和动画的缩写。

统计数据	en⇒de	en⇒fr	en⇒zh	ko⇒zh	中文⇒英文	中文⇒泰文
跨语言族	✗	✗	✓	✓	✓	✓
跨语言分支	✗	✓	✓	✓	✓	✓
时期	2013年-2024	2015年-2024	2020-2024	2017-2024	2021-2024	2021-2024
# 项目	210	207	161	139	225	218
项目类型	电影,电视,动画	电影,电视,动画	电影,电视,动画	TV	电影,电视,纪录片	电影,电视,纪录片
# 行数	1.87M	1.72M	1.02M	1.54M	1.26M	1.21M
总长度（小时）	1156	1033	852	802	501	479
# 平均源语言词元	7.82	7.67	7.77	9.87	6.19	6.08
# 平均目标语言词元	10.83	10.64	6.41	6.51	8.17	21.11

A.2 双语平行语料库构建

通常，同一节目的多语言字幕并非逐行对应。这主要是由语言差异造成的，在人工翻译过程中，原始字幕行可能会合并或拆分。此外，字幕文件中还可能包含场景标题和注释等噪声。为了实现多语言字幕的自动逐行对齐，我们设计了一种利用字幕时间信息的字幕对齐算法。

对于源语言和目标语言的行集合 $\mathcal{L}_{\text{src}}$ 和 $\mathcal{L}_{\text{tgt}}$，我们首先确定行边距 $\mathrm{M} = \mathrm{abs}(|\mathcal{L}_{\text{src}}| - |\mathcal{L}_{\text{tgt}}|)$。然后，对于 $\mathcal{L}_{\text{src}}$中的每一行 ℓ_{src}，我们在 $\mathcal{L}_{\text{tgt}}$ 中对应索引附近大小为 2M 的窗口内搜索一行，其开始时间与 ℓ_{src} 的开始时间相差不超过 0.7 秒，并将其作为 ℓ_{tgt}。选择 0.7 秒作为阈值的原因是，不同语言中最短的句子（如“你好”）的时长通常约为 0.7 秒。通过

这一过程，我们可以从节目中的双语字幕自动收集训练数据集 $\mathbb{D} \equiv \{(\ell_{\text{src}}, \ell_{\text{tgt}}) \in \mathcal{L}_{\text{src}} \times \mathcal{L}_{\text{tgt}}\}$。这一过程的复杂度为 $O(N)$。

B 实验设置的补充细节

本节介绍本文实验所采用的主要设置。除部分采用特殊超参数的实验外，我们均采用相同的超参数，以确保实验比较的一致性和公平性。

B.1 主要设置

我们主要使用 PyTorch[1], vllm[2], 和 Transformers[3] 库来实现我们的方法，并使用 DeepSpeed[4] 进行多 GPU 并行训练。所有实验均在一台配备八块 A800 80GB SXM GPU 的机器上进行。每个翻译方向的 ALPO 采样阶段约需 1.5 小时，训练阶段则需约 2 小时。为确保比较的公平性和一致性，在主要实验、消融研究和扩展实验中，只要条件允许，我们都保持无关参数相同。所有关键超参数设置均列于表 9。

表 9：实验中的超参数配置

阶段	超参数	值	备注
SFT	n	35	提示中的行数 x
	优化器	AdamW	-
	学习率	1e-6	-
	训练轮次	4	-
	批量大小	96	-
ALPO	k	15	采样规模
	温度	1.0	采样生成超参数
	top-k	40	
	top-p	0.9	
	优化器	AdamW	-
	学习率	1e-6	-
	训练轮次	1	-
	批量大小	96	片段级批量大小

我们重新实现了 VideoDubber 方法，并使用阿里云、DeepSeek、OpenAI 和 Anthropic 提供的付费 API，获取通义千问、DeepSeek、GPT 和 Claude 等模型的实验结果和评估结果。通过上下文学习（ICL）在测试集上从这些模型获取译文时，我们使用附录 F.1 中的提示，并提供一个单样本示例，以确保输出格式规范。我们实验中使用的所有大语言模型均为截至 2025 年 9 月 1 日的最新版本，具体包括：qwen-max-2025-01-25[5], DeepSeek-V3.1[6], gpt-4o-2024-08-06[7], GPT-5[8], 以及 Claude 4/4.1[9]。

为进行翻译质量评估，我们将每个数据集中10%的项目保留为测试集。由于付费API成本与推理时间等限制，我们采用基于抽样的评估，而非全量评估。从这些测试项目中，我们随机选取2,000个字幕片段，每个片段包含35行字幕，从而为每个方向构建约70,000个句子的测试集。最终评估结果为这2,000个片段得分的平均值。

[1]https://github.com/pytorch/pytorch[2]https://github.com/vllm-project/vllm[3] https://github.com/huggingface/transformers[4]
https://github.com/microsoft/DeepSpeed[5]
https://help.aliyun.com/zh/model-studio/what-is-qwen-llm[6]
https://api-docs.deepseek.com/news/news250821[7]
https://platform.openai.com/docs/models[8]https://openai.com/gpt-5[9]
https://docs.anthropic.com/en/api/models-list

B.2 实证调查实验

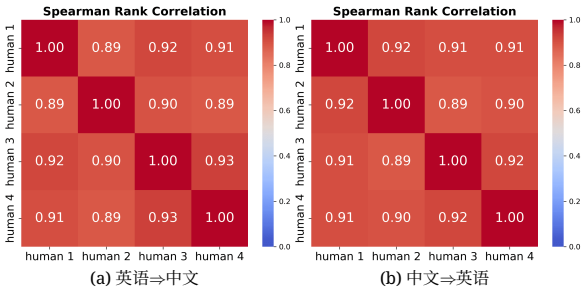
B.2.1 关于 LLM-as-a-Judge 的调查

在en⇒zh、en⇒de、zh⇒en和zh⇒th方向上，我们考察了不同大语言模型 π_e（包括DeepSeek-R1、GPT-5 Thinking、Claude Opus 4.1和Qwen3-14B）与人工评估者在多行字幕多种译文上的评估一致性。具体而言，对于每个方向，每位评估者对500行 $\{s_i \mid i \in [n], 1 \le i \le 500\}$ 在10种不同译文上按0-100分进行评分。这些译文包括人工参考、谷歌翻译、一个SFT模型 π_{sft}，以及多个大语言模型（包括DeepSeek-R1、GPT-5 Thinking、Qwen3-235B-A22B Thinking、GPT-4o、DeepSeek-V3、Claude Sonnet 4和通义千问-Max）。该SFT模型 π_{sft} 使用MuSC 数据集进行训练。评估大语言模型基于上下文学习对译文进行评估；由于行通常与其上下文环境密切相关，我们还提供了 s_i 的上下文行，以便大语言模型参考 s_i的语义环境。提示详见附录 F.2。评分标准如下：

- 原则 1（准确性）：翻译必须确保基本准确性（权重：30%）；
- 原则 2（口语表达得体性）：翻译必须符合人工口语表达的习惯（权重：30%）；
- 原则 3（表现力）：翻译需要生动且富有表现力（权重：40%）。

这里的“权重”一词是为模型提供重要性的指示性参考，而非用于明确的加权操作。

对于英语⇒中文和中文⇒英语这两个方向，人工评估者小组由四名评估者（同四人）组成；英语⇒德语和中文⇒泰语这两个方向则各有两名评估者。所有评估者均为英语、德语或泰语翻译专业的本科生或硕士生，母语为中文。人工评分取各评估者评分的平均值。在图 5 中，我们验证了英语⇒中文和中文⇒英语方向四名评估者之间的一致性，确认了人工评估的可靠性



(a) 英语⇒中文　(b) 中文⇒英语

图 5：人工评估一致性验证。

结果。对于每个评估大语言模型 π_e，我们将译文的命令打乱四次，并计算四次评分的平均值。每位评估者都会对 s_i的多条译文给出分数序列 $\mathcal{E}$。由于不同评估者的评分在 0–100 分范围内通常有所不同，因此我们只关注 $\mathcal{E}$ 的相对排名，而非绝对值。因此，我们评估不同评估者之间的斯皮尔曼等级相关系数 ρ，而不采用组内相关系数等对绝对分数敏感的指标。

B.2.2 回译一致性研究

我们研究了不同领域平行语料库的回译一致性。具体而言，对于某一翻译方向，我们使用 GPT-4o，并通过下方文本框中的简单提示，将目标语言文本回译为源语言。随后，我们计算回译文本与原文之间的 BLEU 和 ChrF++ 分数。分数越低，表示意译程度越高。我们对每个领域评估了 1000 对句子。评估涉及的不同领域数据集包括：

- **OpenSubtitles**（Lison 和 Tiedemann，2016年）是一个包含 60 种语言电影字幕译文的集合。
- **书籍**（Tiedemann，2012年）是文学领域中一系列无版权限制的书籍。
- **bible-uedin**（Christodouloupoulos 和 Steedman，2015年）是一个由《圣经》译文构成的多语言平行语料库。

• **DGT** （Steinberger 等人，2013年）是一个可公开访问的欧盟共同体法多语言翻译记忆库。

• **JRC-Acquis**（Steinberger 等人，2006年）是欧盟立法文本的汇编，包含自20世纪50年代至今编写的精选文件。

• **News-Commentary** （Tiedemann，2012年）是由 WMT 提供的新闻领域平行语料库。

• **ECDC** （Steinberger 等人，2014年）是欧洲疾病预防控制中心的翻译记忆库。

• **EMEA**（Tiedemann，2012年）是由欧洲药品管理局的文件构成的平行语料库。

GPT-4o 的英语⇒德语回译提示示例。

请将以下英文文本逐字翻译成德语，尽可能保持高度逐字对应，不进行任何本地化改写：
起初，地是空虚混沌。渊面黑暗。神的灵运行在水面上。

B.3 人工翻译质量评估

我们对 ALPO 在 en⇒zh 和 zh⇒th 方向上的翻译质量进行了人工评估。每个方向我们都聘请了四名评估员，他们均为英语或泰语翻译专业的本科生或硕士学位持有者，母语均为中文。我们从准确性、自然度和生动性等维度，以及全面性评估方面，将 ALPO 模型与四个基线系统（黄金参考、SFT模型 π_{sft}、GPT-4o 和 DeepSeek-R1）进行了比较。提供给评估员的全面性评估指令详见附录 F.3，其他维度的评估也采用了类似指令。两个方向的多维评估总计包含 32 项评估任务。每位评估员分配了 4 项评估任务。在每项评估任务中，我们向评估员提供了来自 MuSC测试集的 400 个字幕片段的挑战方译文和对照译文，每个片段包含 20 行字幕。评估员需要从两个选项中选出较好的译文，或将其标记为“无显著差异”。为不同评估任务选取的字幕片段均为从测试集中随机抽取的子集。

C 扩展实验

在本节中，我们将展示针对 ALPO 的额外评估实验。

C.1 翻译质量的完整评估

我们在表10中给出了翻译质量评估的完整结果。我们对准确性的评估包括三个大语言模型评估器的结果，以及传统翻译评估模型 XCOMET（Guerreiro 等人，2024）。表10中的结果表明，谷歌翻译在同一语系内的翻译方向上取得了显著表现，例如 en⇒de 和 en⇒fr。然而，它在其他语言上的表现较差，这可能是因为同一语系语言之间的译文相对更为简单。以 MADLAD 为代表的传统序列到序列架构的表现明显不如现代大语言模型，即便是专为字幕翻译设计的 VideoDubber，也并未表现出明显优于 NLLB 和 MADLAD 模型的性能。ALPO 与前沿大语言模型之间的比较得出了若干明确结论：

• 不同的大语言模型评估器通过不同模型对译文进行多维度评估，得出了相一致的排名，进一步验证了第3.1节中的结论。

• 尽管人工翻译的准确性和自然度得分较低，但在生动性维度上表现出色，这可能表明人工译者会结合视频内容，融入更多诠释性翻译。

• 推理模型的生动性得分高于聊天模型，但在其他维度上的得分较低，这表明与聊天模型相比，推理模型更遵循意译指令。这与第 3.3 节的结论一致，因为意译在一定程度上会牺牲准确性。

C.2 翻译示例

表 11 展示了 en⇒zh 和 zh⇒en 方向的翻译示例，其中包括人工翻译、聊天模型 GPT-4o、推理模型 GPT-5 Thinking 以及我们的 ALPO 模型。直观来看，GPT-4o 的译文更偏向直译，未能像其他译文那样通过意译传达原字幕的情感氛围和风格信息。使用 ALPO 训练的模型更符合字幕翻译场景，具体表现如下：

• 译文更口语化，听起来更像人工角色对白，而非正式翻译。

• 措辞和表达方式更加生动，富有表现力。
• 它不会过度追求准确性，而是允许根据上下文进行意译。

此外，表 11 展示了大语言模型翻译的一些局限性。大语言模型翻译仅依赖字幕文本，无法像人工译者那样参考视频语境和场景中的信息。例如，大语言模型将“nature”翻译为“自然”，但根据视频语境，应翻译为“天性”。这一问题超出了偏好优化技术所能解决的范围。因此，我们认为，多模态机器翻译可能是推动字幕翻译研究发展的一个有价值方向。

C.3 扩展消融研究

C.3.1 对齐损失

简化后，ALPO 可以被视为一个兼容各种即插即用偏好优化损失的框架。具体而言，ALPO 损失可以表示为：

$$\mathcal{L}_{\text{alpo}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,\mathcal{S}(x))\sim\mathbb{D}_{\text{alpo}}}\left(\sum_{i=1}^{n} w(s_i)\cdot\mathcal{L}_{\text{po}}(s_i)\right), \tag{7}$$

其中，$\mathcal{L}_{\text{po}}$ 可以采用多种偏好优化损失，例如 DPO（Rafailov 等，2024年）、SimPO（Meng 等，2024年）和 GRPO（Shao 等人，2024年）。对于基于优势的群体相对策略优化损失，我们将其计算为：

$$\mathcal{L}_{\text{po}}(s_i) = \frac{1}{|\mathcal{T}_i|}\sum_j\left[\min\left(\frac{\pi_\theta(t_i^j\mid p_i)}{\pi_{\text{ref}}(t_i^j\mid p_i)}A_j, \text{clip}\left(\frac{\pi_\theta(t_i^j\mid p_i)}{\pi_{\text{ref}}(t_i^j\mid p_i)}, 1-\varepsilon, 1+\varepsilon\right)A_j\right) - \eta\text{KL}(\pi_\theta||\pi_{\text{ref}})\right], \tag{8}$$

$$A_j = \frac{\mathcal{E}_i^j - \text{mean}(\mathcal{E}_i)}{\text{std}(\mathcal{E}_i)}, \quad \text{KL}(\pi_\theta||\pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(t_i^j\mid p_i)}{\pi_\theta(t_i^j\mid p_i)} - \log\frac{\pi_{\text{ref}}(t_i^j\mid p_i)}{\pi_\theta(t_i^j\mid p_i)} - 1, \tag{9}$$

其中，ε 和 η 是可调超参数，而 A_j 表示根据分数序列 $\mathcal{E}_i$估算出的优势。

我们在表12中比较了 ALPO 使用不同偏好对齐损失 $\mathcal{L}_{\text{po}}$ 时的表现。实验结果表明，所有损失变体均显著优于监督微调模型，验证了 ALPO 广泛的兼容性。DPO 损失在多个翻译方向上表现稳定且出色。尽管 GRPO 有效利用了所有采样输出，而非仅关注偏好对，但它仅在 en⇒zh 方向上取得最佳表现。

Table 10：完整的质量评估结果。排名第 1 和第 2 的结果分别标记为 **蓝色** 和 **色**，resp别地。

en⇒de

模型	训练	准确性 XCOMET	准确性 DeepSeek	准确性 Claude	准确性 GPT-5	准确性 Avg.	自然度 DeepSeek	自然度 Claude	自然度 GPT-5	自然度 Avg.	生动性 DeepSeek	生动性 Claude	生动性 GPT-5	生动性 Avg.
黄金参考	人工	81.4	83.4	88.6	85.6	84.8	83.6	85.5	82.2	83.8	64.4	**77.1**	**77.7**	**73.1**
VideoDubber	ST	43.2	46.1	55.1	21.4	41.5	44.7	40.2	21.6	35.5	42.2	47.1	33.7	41.0
NLLB-3.3B	ST	70.3	73.5	88.1	71.4	75.8	72.2	75.2	63.0	70.1	58.2	60.4	58.9	59.2
MADLAD-10B	ST	68.9	70.0	87.3	66.4	73.2	68.4	71.1	57.2	65.6	55.2	55.7	52.6	54.5
谷歌翻译	ST	88.3	89.9	89.9	91.5	89.9	80.4	81.7	79.8	80.6	60.9	64.2	62.8	62.6
GPT-4o	ICL（C）	91.2	96.8	96.5	91.8	94.1	89.1	88.7	82.2	86.7	56.2	71.3	73.3	66.9
通义千问-Max	ICL（C）	**91.7**	**97.8**	**97.9**	94.7	**95.5**	**89.6**	**90.1**	**87.8**	**89.2**	58.0	74.0	74.7	68.9
DeepSeek-V3.1	ICL（C）	90.2	97.3	96.9	**94.9**	94.8	**89.8**	**89.9**	**87.9**	**89.2**	56.5	71.4	74.2	67.4
DeepSeek-R1	ICL（R）	89.3	94.4	94.4	93.0	92.8	88.0	88.7	85.5	87.4	61.1	74.9	74.1	70.0
GPT-5	ICL（R）	90.7	95.7	94.6	93.4	93.6	89.1	89.3	87.5	88.6	**64.8**	76.5	76.9	72.7
QwenZ5-14B	SFT	84.7	87.6	91.1	86.7	87.5	83.9	86.3	80.6	83.6	56.0	68.9	68.2	64.4
Qwen2.5-14B	**ALPO**	**91.9**	**98.0**	**97.0**	**94.8**	**95.4**	88.9	89.3	87.1	88.4	**68.3**	**78.0**	**78.1**	**74.8**

en⇒fr

模型	训练	准确性 XCOMET	准确性 DeepSeek	准确性 Claude	准确性 GPT-5	准确性 Avg.	自然度 DeepSeek	自然度 Claude	自然度 GPT-5	自然度 Avg.	生动性 DeepSeek	生动性 Claude	生动性 GPT-5	生动性 Avg.
黄金参考	人工	76.4	84.4	89.0	84.1	83.5	85.5	86.8	82.7	85.0	68.1	78.6	77.6	74.8
VideoDubber	ST	42.3	46.9	62.1	41.7	48.2	48.1	52.5	41.4	47.3	47.4	53.4	44.8	48.5
NLLB-3.3B	ST	62.2	76.9	89.6	78.7	76.9	75.0	77.9	70.8	74.6	61.4	63.3	60.7	61.8
MADLAD-10B	ST	58.1	75.6	87.5	73.3	73.6	70.3	71.4	61.7	67.8	55.7	59.4	57.1	57.4
谷歌翻译	ST	81.4	93.5	95.9	**96.7**	91.9	84.5	85.8	84.2	84.8	64.3	64.6	64.0	64.3
GPT-4o	ICL（C）	86.4	**97.7**	96.3	92.6	93.2	89.8	90.3	86.4	88.8	59.0	74.9	74.7	69.5
通义千问-Max	ICL（C）	86.1	97.5	**97.3**	95.6	**94.1**	**90.4**	90.4	89.0	89.9	61.1	76.8	76.8	71.6
DeepSeek-V3.1	ICL（C）	**87.6**	97.2	**98.9**	**95.9**	**94.9**	**90.8**	90.6	**90.7**	**90.7**	61.7	77.3	77.9	72.3
DeepSeek-R1	ICL（R）	**87.1**	96.7	95.3	95.1	93.6	89.7	**90.8**	89.6	90.0	64.5	77.3	78.3	73.4
GPT-5	ICL（R）	83.8	96.4	95.0	93.8	92.2	90.1	90.8	**90.4**	**90.4**	**68.4**	**79.3**	**79.8**	**75.8**
QwenZ5-14B	SFT	77.7	89.7	90.6	86.7	86.2	85.9	86.8	83.6	85.4	57.4	71.7	74.0	67.7
Qwen2.5-14B	**ALPO**	86.5	**97.9**	96.7	95.4	94.1	88.7	**91.5**	87.4	89.2	**73.7**	**82.0**	**80.7**	**78.8**

英⇒中

模型	训练	准确性 XCOMET	准确性 DeepSeek	准确性 Claude	准确性 GPT-5	准确性 Avg.	自然度 DeepSeek	自然度 Claude	自然度 GPT-5	自然度 Avg.	生动性 DeepSeek	生动性 Claude	生动性 GPT-5	生动性 Avg.
黄金参考	人工	79.0	84.3	88.5	82.6	83.6	81.6	85.7	80.6	82.6	68.9	69.7	**76.0**	**71.5**
VideoDubber	ST	40.9	47.5	64.1	35.0	46.9	52.4	61.3	42.0	51.9	49.7	52.2	47.1	49.7
NLLB-3.3B	ST	61.9	65.2	70.4	48.2	61.4	59.5	56.8	45.6	54.0	46.3	45.2	39.5	43.7
MADLAD-10B	ST	53.5	61.0	79.5	44.7	59.7	58.3	65.5	42.7	55.5	47.4	46.7	44.8	46.3
谷歌翻译	ST	79.3	81.4	**96.1**	80.0	84.2	79.4	85.1	74.6	79.7	52.3	51.8	59.0	54.4
GPT-4o	ICL（C）	84.1	92.1	92.6	88.6	89.3	83.0	84.2	79.6	82.3	56.3	57.7	65.3	59.8
通义千问-Max	ICL（C）	**88.3**	95.3	93.5	90.3	**91.9**	85.8	85.8	81.6	84.4	59.1	61.7	63.1	61.3
DeepSeek-V3.1	ICL（C）	85.9	**95.6**	**93.6**	89.5	91.2	86.9	**86.2**	82.8	85.3	59.5	61.8	69.2	63.5
DeepSeek-R1	ICL（R）	86.9	92.3	92.9	90.0	90.5	**87.2**	86.0	**83.9**	**85.7**	68.1	**70.0**	74.3	70.8
GPT-5	ICL（R）	**88.4**	**96.7**	93.6	**90.8**	**92.4**	**88.4**	**87.1**	**85.4**	**87.0**	**69.2**	69.9	74.3	71.1
QwenZ5-14B	SFT	82.4	89.9	90.1	83.4	86.4	83.3	83.3	79.5	82.0	57.0	56.0	64.2	59.1
Qwen2.5-14B	**ALPO**	84.8	93.9	93.1	**90.8**	90.6	84.2	85.0	83.8	84.3	**73.6**	**76.9**	**79.2**	**76.6**

ko⇒zh

模型	训练	准确性 XCOMET	准确性 DeepSeek	准确性 Claude	准确性 GPT-5	准确性 Avg.	自然度 DeepSeek	自然度 Claude	自然度 GPT-5	自然度 Avg.	生动性 DeepSeek	生动性 Claude	生动性 GPT-5	生动性 Avg.
黄金参考	人工	68.6	80.4	88.4	74.6	78.0	78.8	79.9	74.7	77.8	**65.6**	58.4	**73.5**	**65.8**
VideoDubber	ST	35.4	47.0	56.8	19.3	39.6	48.9	58.2	28.5	45.2	47.4	56.0	41.1	48.2
NLLB-3.3B	ST	30.9	39.8	40.1	21.8	33.1	30.3	29.5	18.5	26.1	27.6	30.6	17.9	25.4
MADLAD-10B	ST	37.2	47.6	66.1	28.8	44.9	44.7	57.3	26.8	42.9	44.4	55.6	40.2	46.7
谷歌翻译	ST	42.4	50.6	76.9	49.6	54.9	49.5	61.8	47.2	52.8	46.3	57.3	52.4	52.0
GPT-4o	ICL（C）	70.5	87.5	86.5	75.4	80.0	81.7	81.7	76.4	79.9	55.5	51.8	67.0	58.1
通义千问-Max	ICL（C）	74.5	89.7	**88.8**	**81.7**	83.7	83.4	**83.7**	**80.5**	82.5	57.1	57.6	70.7	61.8
DeepSeek-V3.1	ICL（C）	73.5	**91.9**	87.7	79.1	83.1	**84.7**	82.2	79.8	82.2	55.6	49.6	66.4	57.2
DeepSeek-R1	ICL（R）	71.1	86.0	86.4	75.9	79.8	84.0	82.3	78.6	81.6	64.0	**60.0**	72.7	65.6
GPT-5	ICL（R）	**76.3**	**93.5**	88.7	79.5	**84.5**	**85.1**	83.2	79.6	**82.6**	63.1	59.3	72.5	65.0
QwenZ5-14B	SFT	72.9	84.8	88.1	77.7	80.9	78.5	78.7	71.2	76.1	53.8	47.0	61.0	53.9
Qwen2.5-14B	**ALPO**	**76.0**	89.8	**89.0**	**82.3**	**84.3**	83.6	**84.7**	**81.7**	**83.3**	**67.6**	**67.5**	**76.3**	**70.5**

zh⇒en

模型	训练	准确性 XCOMET	准确性 DeepSeek	准确性 Claude	准确性 GPT-5	准确性 Avg.	自然度 DeepSeek	自然度 Claude	自然度 GPT-5	自然度 Avg.	生动性 DeepSeek	生动性 Claude	生动性 GPT-5	生动性 Avg.
黄金参考	人工	78.0	87.8	86.6	79.4	83.0	82.0	81.9	77.1	80.3	71.9	71.1	76.9	73.3
VideoDubber	ST	54.1	54.3	66.5	39.5	53.6	59.2	60.5	44.6	54.8	53.9	52.8	43.6	50.1
NLLB-3.3B	ST	35.9	30.8	36.8	12.7	29.1	26.4	26.1	12.5	21.7	26.4	24.4	11.5	20.8
MADLAD-10B	ST	41.6	49.6	56.0	33.3	45.1	45.7	40.9	30.2	38.9	43.5	38.2	31.2	37.6
谷歌翻译	ST	77.2	75.6	88.1	78.2	79.8	70.0	67.6	61.3	66.3	56.2	45.4	49.1	50.2
GPT-4o	ICL（C）	84.8	92.3	90.3	86.7	88.5	85.2	83.9	79.9	83.0	61.8	63.4	68.6	64.6
通义千问-Max	ICL（C）	**86.4**	**93.2**	**92.6**	87.9	**90.0**	**87.0**	85.2	82.7	85.0	63.7	67.2	69.5	66.8
DeepSeek-V3.1	ICL（C）	**85.3**	**92.9**	**91.4**	88.3	**89.5**	85.8	84.9	81.5	84.1	59.0	63.8	66.2	63.0
DeepSeek-R1	ICL（R）	83.2	91.6	89.6	**89.8**	88.5	85.6	83.9	**87.4**	85.6	69.5	71.3	**79.7**	73.5
GPT-5	ICL（R）	84.9	92.0	90.9	**88.5**	89.1	**87.2**	**86.8**	84.3	**86.1**	**72.7**	**75.3**	77.5	**75.2**
QwenZ5-14B	SFT	80.4	88.7	88.2	83.5	85.2	81.9	82.0	76.3	80.1	51.5	52.3	60.7	54.8
Qwen2.5-14B	**ALPO**	85.1	90.9	89.8	87.6	88.3	86.1	**87.1**	**87.2**	**86.8**	**79.2**	**83.0**	**82.8**	**81.7**

zh⇒th

模型	训练	准确性 XCOMET	准确性 DeepSeek	准确性 Claude	准确性 GPT-5	准确性 Avg.	自然度 DeepSeek	自然度 Claude	自然度 GPT-5	自然度 Avg.	生动性 DeepSeek	生动性 Claude	生动性 GPT-5	生动性 Avg.
黄金参考	人工	70.5	77.9	78.0	80.0	76.6	74.3	74.7	76.3	75.1	60.8	67.5	70.6	66.3
VideoDubber	ST	40.2	38.2	42.5	15.7	34.1	39.0	48.7	17.0	34.9	43.7	49.8	31.0	41.5
NLLB-3.3B	ST	41.8	50.2	59.8	18.7	42.6	48.9	41.1	11.7	33.9	49.3	41.7	30.5	40.5
MADLAD-10B	ST	45.6	51.4	69.0	25.7	47.9	54.7	64.7	33.0	50.8	52.7	58.1	42.2	51.0
谷歌翻译	ST	52.5	50.1	71.4	46.7	55.2	53.6	65.9	49.2	56.2	53.5	57.6	52.5	54.5
GPT-4o	ICL（C）	80.9	90.8	89.6	90.7	88.0	**85.1**	84.7	83.5	84.4	62.6	67.1	73.9	67.9
通义千问-Max	ICL（C）	84.3	**95.1**	**91.4**	**94.5**	**91.3**	**85.9**	**86.1**	**85.5**	**85.8**	63.8	70.0	73.6	69.1
DeepSeek-V3.1	ICL（C）	**84.8**	92.6	89.3	93.1	89.9	85.0	85.2	83.5	84.6	61.8	66.6	72.8	67.1
DeepSeek-R1	ICL（R）	82.2	89.2	88.1	90.8	87.6	83.5	84.4	84.0	84.0	64.4	72.3	76.2	71.0
GPT-5	ICL（R）	83.3	92.9	88.1	90.5	88.7	85.1	82.4	84.1	83.9	**67.3**	**73.7**	**78.1**	**73.0**
QwenZ5-14B	SFT	81.7	90.0	88.8	88.7	87.3	82.6	83.8	81.4	82.6	59.5	66.3	72.2	66.0
Qwen2.5-14B	**ALPO**	**85.0**	**95.0**	**92.7**	**95.1**	**91.9**	84.3	**85.3**	**84.4**	**84.7**	**69.0**	**75.4**	**78.1**	**74.2**

表 11：英语⇒中文 与 中文⇒英语的多条翻译示例。高质量译文和质量不足的译文分别以**蓝色**和**橙色**标示。

源文本	翻译			
	人工	GPT-4o	GPT-5	ALPO（我们的方法）
皮尔特沃夫的建立是为了逃离法师们的好战行径。	当初建立皮尔特沃夫就是为了避开那群好战的魔法师	建立皮尔特沃夫正是为了逃离法师的战争	皮尔特沃夫是为了逃避法师的战争狂澜而建立的	皮尔特沃夫的创立是为了逃离术士们的
秘术是我们世界的诅咒。我的族人几乎因此被毁灭。	秘术对我们的世界来说就是诅咒，我的族人几乎因此而灭绝	秘术是我们世界的诅咒我的族群几乎被它毁灭	秘术是我们世界的诅咒我的族人几乎因此灭绝	秘术是世间的诅咒 我的族人曾险些被它
霍斯卡尔议员和萨罗议员。他们不是互相讨厌吗？	霍斯卡尔议员和萨罗议员他俩不是 吗	议员们霍斯克和萨罗他们不是互相讨厌吗	霍斯克尔议员和萨罗议员他俩不是水火不容吗	霍斯卡尔议员和萨罗议员他们不是 吗
他的一生，他一直追逐着一个不可能实现的梦想。	他这辈子都在追逐一个不可能实现的梦想	他一生都在追逐一个不可能的梦想	他这辈子都在追逐不切实际的幻想	他一生都在追逐一个 可及的梦
那是梦想之船，驶向其他所有人。对我而言，它是一奴隶船。	它是所有人心目中的梦想之船对我来说只是一条贩奴船	对其他人来说 这是一艘梦想之船 对我而言 它是一艘奴隶船	在别人眼中，这是艘dream之船，但对我来说，它却是艘奴隶船	对旁人来说，这艘船 dream，而于我而言，它不过是艘奴隶船
我本来会跳下船去，但是道森先生救了我。	我差点儿掉下去是道森先生救了我	如果不是道森先生救了我我早就掉下水了	要不是这位道森先生救我我差点就掉下去了	要不是道森先生及时相救我差点
大自然让我们难以接受改变但幸运的是，我们有能力改变我们的天性。	天性会让我们排斥异变，但好在，我们有能力改变自己的天性	天性让我们难以接受变化，但幸运的是，我们有能力改变我们的本性	自然让我们无法适应剧变但幸运的是 我们拥有改变天性的能力	自然造物使我们抗拒改变但我们幸而拥有改变的天性的能力
你让我想去看那世界不只是黑与白。	你让我超越黑白分明的视角看世界	你让我想要从不是非黑即白的角度看世界	你让我学会用更丰富的眼光看世界	你让我渴望用 缤纷的色彩 而不只是黑白两色来观照世界
别害怕。想想自己是一只毛毛虫，即将蜕变成蝴蝶。	别害怕，想象自己即将变得丰满。即将化蝶的毛毛虫	别害怕，想象一下自己是一只即将 的毛毛虫	别害怕，想象自己是一只即将化蝶的毛毛虫	别害怕，想象自己是一只即将 的毛毛虫
只要我们保护彼此我们就能永葆不朽……而且永远年轻。	只要我们互相保护，就会永远年轻	只要我们保护彼此，就可以长生不老，青春永驻	只要互相守护 就能永葆青春……与不朽	只要我们彼此守护，就能 生不老，
你们两个觉得自己更胜一筹比我们强，但说真的，你是个假清高的人和一头猪。	你们俩以为自己比我们强但你们只是猪猡	你们俩以为自己比我们好但事实上 你们不过就是一个牛鼻子老学究和一个小猪	你们俩自以为高人一等，其实一个假清高 一个真肥猪	你们俩自以为高人一等，其实啊，你俩一个 一个
有一天你风光无限……第二天，你就成了小丑。	今日 明朝小丑相见	有一天你在上面…下一天你就是个小丑	今天你风光无限…明天就沦为小丑	你今天还 明日就沦为小丑一枚
哥谭市喜欢东山再起的故事。	哥谭最喜欢的就是 归	Gotham City 的故事	Gotham City最爱看逆袭故事	Gotham residents 身的戏码
你是……的实际领导人 议会的实际领导人。人们会注意你的行踪。	你才是议会真正的领袖 你的行踪备受关注	你才是议会的实际领导人 人们	身为议会实际上的话事人 行踪自然会受到关注	You are the de facto leader of the parliament Naturally, every move you make attracts attention
At that time, I had a strong premonition I knew I would definitely make that shot	在那一刻，我有着非常强烈的 直觉。 我当时觉得自己能够投进这一球。	那一刻，我我有种强烈的感觉我肯定会投进那一球。	当时，我只是有这种强烈的感觉——那一球，你知道，我当时觉得自己能做到。	那一刻，我有种强烈的预感， 觉得自己能投进那一球。
霜华清光 凝息止气	明亮的霜驱散一切杂质。封住她的 aura.	Frost的纯 平静的呼吸和能量。	霜光汇聚，气息静止。	Frost Illume, 静住呼吸，平息空气。
人生中有许多重要决定都是在这种稀里糊涂的时候马上做出的	有许多重要决定是我们在生活中不得不做的立即做出当情况仍然 模糊不	许多重要生活中的重要决定必须做出在这些混乱 时刻。	许多关键的人生决策必须恰恰在这些混乱 立即。	许多重要的生活中的决定必须当场 在困惑时刻困惑时刻和茫
现在恐怕还不是谈情说爱的时候	恐怕现在不是……的时候你们两个谈情说爱 卿卿我我。	现在大概不是谈浪漫的时间。	大概还不是时间去谈	恐怕现在还不是合适的时间来 卿 呢。
你我皆是被命运捉弄之人为何不与我一起颠覆了这世间	你我二人都被命运玩弄。你为何不加入我，一起 这个世界?	你和我都是命运的玩物。何不与我一同 推翻这个世界?	我们都是命运的受害者残酷。为何不 与我一同推翻这个世界?	我们俩都是随心摆布的棋子命运的棋子。为何不和我联手一起让这个世界天翻地覆?
只可惜这和煦的暖阳以后再也看不到了	真可惜，我将无法再感受到温暖的阳光了。	真遗憾这温暖的阳光，我将再也见不到了。	只是可惜这温暖、柔和的 阳光——我永远不会再见到它了。	真可惜，我再也无法沐浴 存温暖的阳光下了……
你别看我赌运差但我看人的眼光可是很准的	我可能不太走运说到赌博，但我很擅长判断角色。	别评判我凭我糟糕的运气在赌博，但我的识人眼光相当敏锐。	不要以……来评判我我糟糕的赌博运气——我的识人能力非常敏锐。 锋利无比。	就因为我不是幸运的赌徒，并不代表我不擅长看人方面。我的直觉一向很准。
青山常在，绿水长流	世界将会一如既往。	山是 青的，而水流长。	青山永存，碧水流淌。	愿山巍然耸立，水流不息。
她为了保护你被千夫所指因为担心你而食不下咽	她正面临一场 因试图保护你而遭受强烈反对。而且她太担心你了，吃不下东西。	她把所有的为保护你而责备你，她忧心忡忡，吃不下东西。	她忍 嘲讽来保护你，失去了食欲担心你。	她因为保护你 而遭到诋毁，她还因此没了食欲，都是因为担心你。
她一定有三头六臂	我敢肯定她有很多手臂和腿。	她一定有三个头和六条手臂。	她一定有三个头和六条手臂之类的吧。	她一定是个 女超人。
青丘军队全部停止前行原地扎寨 不得违令	所有青丘军队都必须停止前进并扎营在你们所在之处。这道命令必须必须服从。	青丘军必须停止并立即扎营，不得 允许。	所有青丘军队停止进军，立即扎营。 禁止违令。	青丘的军队，停止进军 立刻。 设置扎营。不许 违令。
人间的桂花开了	人间的桂花 人间正在盛开。	人间 正在盛开于人间。	人间的桂花人间是盛开。	桂花在人间盛开在人间。

表12：影响 $\mathcal{L}_{\text{po}}$。第1和第2名的结果分别以**蓝色**和**橙色**表示。

方法	$\mathcal{L}_{\text{po}}$	英语⇒德语			英语⇒中文			中文⇒英语		
		准确性	自然度	生动性	准确性	自然度	生动性	准确性	自然度	生动性
SFT	-	87.7	83.4	64.4	86.5	82.1	59.2	85.2	80.1	54.9
ALPO	DPO	**95.2**	**88.3**	**74.8**	90.6	**84.2**	**76.6**	**88.3**	**86.8**	**81.6**
	SimPO	93.2	87.4	73.9	**90.7**	83.9	74.3	86.2	84.2	80.9
	GRPO	**94.6**	**87.9**	**74.3**	**91.2**	**85.0**	**77.1**	**87.2**	**84.9**	**81.3**

C.3.2 采样规模

ALPO 的采样规模直接影响采样译文的多样性。在图 6 中，我们考察了采样规模 k 对 zh⇒en 方向翻译质量的影响。结果表明，随着 k 增大，多个维度上的翻译质量均有所提升，其中生动性的提升最为显著，并在 $k = 12$左右趋于稳定。ALPO 训练过程中，自适应超参数 $w(s_i)$ 和 β_i 的计算取决于采样译文的数量及其质量变化。较大的采样规模可直接提升训练译文的质量，并带来更合适的 $w(s_i)$ 和 β_i。由于 ALPO 损失取决于选定译文与拒绝译文的相对质量，因此采样规模足够大时，性能便会趋于稳定。基于这些结果，我们在其他所有实验中均设置为 $k = 15$ 。

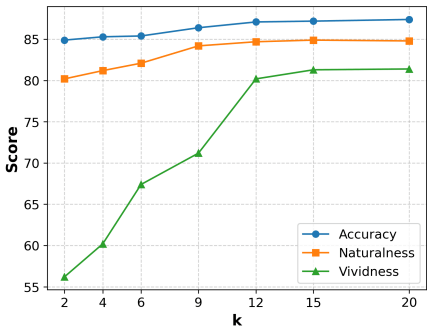


图 6：采样规模对翻译质量的影响。

C.3.3 模型规模

在图7中，我们研究了 ALPO 的骨干模型规模对翻译质量的影响。我们采用从 1.5B 到 72B 不等的 Qwen2.5系列模型作为 ALPO 的基座模型，评估 en⇒zh 和 zh⇒en 两个方向的翻译性能。实验结果表明，随着模型规模增大，各维度的分数均持续提升，但提升幅度逐渐减小。14B 或 32B 模型通常能达到与 DeepSeek-R1 相当的性能，这进一步验证了 ALPO 在提升翻译质量方面的重要作用。ALPO 采用完全自主的方式对对齐数据进行偏好标注，从而能够以较低成本训练高质量的字幕翻译模型。总体而言，14B基座模型在成本和性能之间实现了较好的平衡。不过，由于字幕翻译属于离线任务，若追求 SOTA 性能，也可以采用更大规模的模型。

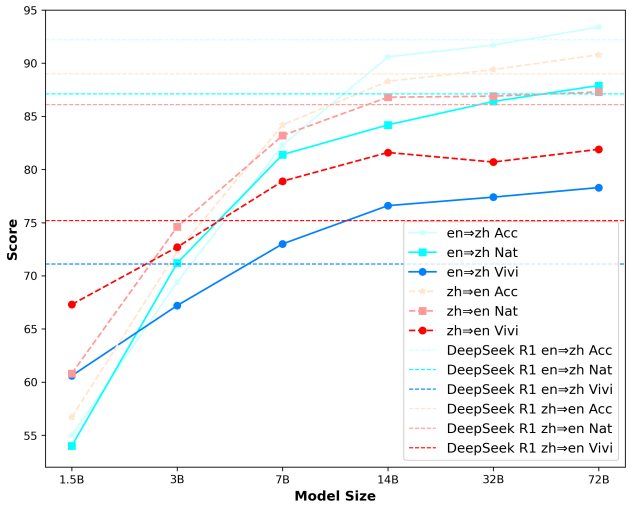


图7：模型规模对翻译质量的影响。

C.3.4 数据量

我们在图 8 中展示了不同训练步数下的翻译质量分数，以研究训练数据规模的影响。结果表明，随着训练数据量增加，模型准确率先升高后下降，自然度略有提升，而生动性则显著提高，但增幅逐渐减缓。尽管增加训练数据可提高生动性评分，我们观察到模型输出中出现了明显的幻觉现象，这解释了准确性下降的原因。全面分析表明，训练 75 步（对应 7,000 个提示，批量大小为 96）可实现最佳平衡，因此我们将其作为所有实验的默认配置。

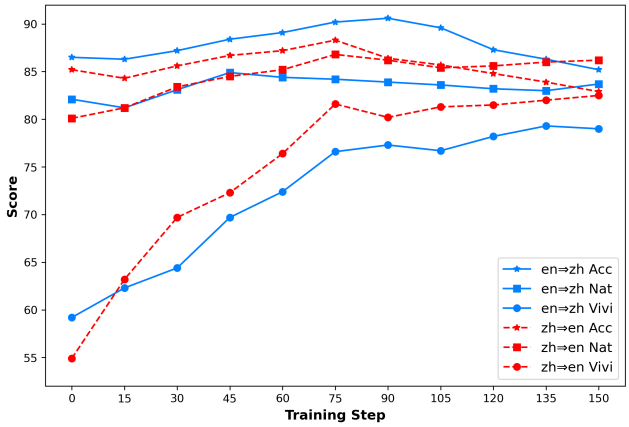


图 8：训练数据量对翻译质量的影响。

C.4 进一步探索

C.4.1 原始 DPO 训练

将富有表现力的字幕翻译大语言模型的训练视为一个偏好优化问题，我们希望探索**原始 DPO 能否直接解决这一问题**。受此探究启发，我们设计并开展了相关实验。ALPO 通过其分段采样策略和细粒度对齐损失，在翻译质量上取得了显著提升，我们旨在验证这两个组成部分的影响。遵循标准 DPO 训练流程（Rafailov 等，2024年），我们在 SFT模型 π_{sft}上进行后训练。具体而言，我们采用粗粒度或细粒度采样，为对齐数据集中的每个样本 $x \in \mathbb{D}_{\text{alpo}}$ 生成一个选定响应 $y^{(c)}$ 和一个拒绝响应 $y^{(r)}$，随后使用标准 DPO 损失优化策略模型。

粗粒度和细粒度采样流程如算法 2 和算法 3 所示。在粗粒度采样方法中，我们针对给定提示 x 直接采样 k 个完整响应。随后，使用 Qwen2-14B-Instruct 模型为这 k 个响应中的所有字幕行评分。计算每个响应中 n 个字幕行的得分总和，并将得分最高的响应选为选定响应 $y^{(c)}$，得分最低的响应选为拒绝响应 $y^{(r)}$。对于细粒度采样，我们针对每个提示 x，执行两轮独立的片段级采样，过程类似算法 1。在这两轮采样中，后续字幕行的采样分别以前缀为现有得分最高或最低的字幕行，最终得到片段级采样的选定响应 $y^{(c)}$ 和拒绝响应 $y^{(r)}$。

C.4.2 基于优势的近端策略优化训练

在文学研究中，人们普遍认为，在偏好优化技术中，RLHF（主要基于近端策略优化）尽管流程复杂，但其表现优于 DPO（Ouyang 等，2022年；Bender 等人，2021年）。这促使我们研究**PPO 是否能够取得优于 ALPO 的性能**。为此，我们设计并开展了相关实验。我们采用

算法 2原始 DPO 的粗粒度采样。**输入：** SFT模型 π_{sft}、评估大语言模型 π_{e}、对齐数据集 $\mathbb{D}_{\text{alpo}}$、采样数量 k。**输出：** 采样响应对集合 $\mathcal{S}(x)$。1：**对于**任意 $x \in \mathbb{D}_{\text{alpo}}$ **执行**// 遍历对齐数据集 $\mathbb{D}_{\text{alpo}}$。2：**对于** $i = 1$ 至 k **执行**// 采样多个候选响应。3：采样 $\pi_{\text{sft}}(y \mid x)$。4：**结束循环**5：计算候选集 $\left\{y^i | i = 1, 2, \ldots, k\right\}$ 中每一行 y^i 的 π_{e} 分数 $\mathcal{E}$之和。6：选出 chosen $y^{(\text{c})}$ 和 rejected $y^{(\text{r})}$。7：**结束循环**8：**返回** $\mathcal{S}(x) \equiv \left\{y^{(\text{c})}, y^{(\text{r})}\right\}$。

算法 3 面向原始 DPO 的细粒度采样。输入：SFT模型 π_{sft}、评估大语言模型 π_{e}、对齐数据集 $\mathbb{D}_{\text{alpo}}$、采样数量 k。输出：采样响应对集合 $\mathcal{S}(x)$。1：对任意 $x \in \mathbb{D}_{\text{alpo}}$ 执行 //遍历对齐数据集 $\mathbb{D}_{\text{alpo}}$。2：对于 $i = 1$到 n 执行 // 用于获取 $y^{(\text{c})}$ 的第一次采样循环。3：对于 $j = 1$ 到 k 执行 4：采样 $\pi_{\text{sft}}(t_i^j \mid x, t_1^{(\text{c})}, \ldots, t_{i-1}^{(\text{c})})$。5：结束循环 6：对候选集 $\{t_i^j \mid j = 1, 2, \ldots, k\}$ 去重。7：将人工参考 t_i^0 加入候选集，得到 $\{t_i^j \mid j = 0, 1, \ldots\}$。8：用 π_{e} 度量 $\{t_i^j \mid j = 0, 1, \ldots\}$，得到分数序列 $\mathcal{E}_i$。9：选择选中译文 $t_i^{(\text{c})}$（从 $\mathcal{E}_i$ 的前 3 名中随机选取）。10：结束循环 11：拼接 $\{t_i^{(\text{c})} | i = 1, 2, \ldots, n\}$ 得到 $y^{(\text{c})}$。12：对于 $i = 1$到 n执行 //用于获取 $y^{(\text{r})}$ 的第二次采样循环。13：对于 $j = 1$到 k 执行 14：采样 $\pi_{\text{sft}}(t_i^j \mid x, t_1^{(\text{r})}, \ldots, t_{i-1}^{(\text{r})})$。15：结束循环 16：对候选集 $\{t_i^j \mid j = 1, 2, \ldots, k\}$ 去重。17：将人工参考 t_i^0加入候选集，得到 $\{t_i^j \mid j = 0, 1, \ldots\}$。18：用 π_{e} 度量 $\{t_i^j \mid j = 0, 1, \ldots\}$，得到分数序列 $\mathcal{E}_i$。19：选择拒绝译文 $t_i^{(\text{r})}$（从 $\mathcal{E}_i$ 的后 3 名中随机选取）。20：结束循环 21：拼接 $\{t_i^{(\text{r})} | i = 1, 2, \ldots, n\}$ 得到 $y^{(\text{r})}$。22：结束循环 23：返回 $\mathcal{S}(x) \equiv \{y^{(\text{c})}, y^{(\text{r})}\}$。

一种基于优势的近端策略优化训练流程（Zheng 等，2023b）。具体而言，我们通过以下步骤实现近端策略优化训练过程：

1. 轨迹采样 — 使用算法 4 为每个输入 $x \in \mathbb{D}_{\text{alpo}}$ 采样一条轨迹 τ。2. 计算回报与优势 — 使用价值网络 V_ϕ（基座模型：Qwen2.5-7B-Instruct（Yang 等，2024a））对 τ 执行广义优势估计（GAE）：

$$\begin{aligned} \delta_i &= -\mathcal{E}_i + \gamma V_\phi(p_{i+1}) - V_\phi(p_i), \\ A_i &= \sum_{l=0}^{n-i-1} (\gamma\lambda)^l \delta_{i+l}. \end{aligned} \tag{10}$$

3. 更新策略网络 — 令 π_{old} 表示采样期间固定的旧策略。最小化裁剪损失函数（下文省略 π_θ 与 π_{old} 之间的 KL 散度约束）：

$$\mathcal{L}_{\text{clip}}(\theta) = -\mathbb{E}_{x\sim\mathbb{D}_{\text{alpo}}}\left[\sum_{i=1}^{n}\min\left(\frac{\pi_\theta(t_i \mid p_i)}{\pi_{\text{old}}(t_i \mid p_i)}A_i, \text{clip}\left(\frac{\pi_\theta(t_i \mid p_i)}{\pi_{\text{old}}(t_i \mid p_i)}, 1-\epsilon, 1+\epsilon\right)A_i\right)\right]. \tag{11}$$

4. 更新价值网络—通过均方误差将 $V_\phi(p_i)$ 拟合至基于 GAE 的估计 $\hat{V}_i = A_i + V_\phi(p_i)$：

$$\mathcal{L}_V(\phi) = \mathbb{E}_{x\sim\mathbb{D}_{\text{alpo}}}\left[\sum_{i=1}^{n}(V_\phi(p_i) - \hat{V}_i)^2\right]. \tag{12}$$

5. 重复多个训练轮次—持续收集新数据，并更新价值网络和 d 策略网络，直至收敛。

算法 4 近端策略优化解的采样。输入：SFT模型 π_{sft}、评估大语言模型 π_{e}、对齐数据集 $\mathbb{D}_{\text{alpo}}$。输出：采样得到的轨迹集 $\mathcal{S}(x) \equiv \{\tau\}$。1：对于任意 $x \in \mathbb{D}_{\text{alpo}}$，执行//遍历 DA 数据集 $\mathbb{D}_{\text{alpo}}$。2：对于 $i = 1$ 到 n，执行// 遍历 x 中的字幕行。3：采样 $\pi_{\text{sft}}(t_i \mid x, t_1, .., t_{i-1})$。4：通过 π_{e} 测量 t_i，得到分数 $\mathcal{E}_i$。5：结束循环。6：获得轨迹 $\tau = \{(p_i, t_i, \mathcal{E}_i) \mid i = 1、2、\cdots、n\}$。7：结束循环。8：返回 $\mathcal{S}(x) \equiv \{\tau\}$。

C.4.3 评估

我们使用 Qwen2.5-14B 模型，评估了在“原始 DPO 训练”和“基于优势的 PPO 训练”配置下 en⇒zh 和 zh⇒en 翻译的性能。此外，我们还比较了两种基于词元级 DPO 的方法：TDPO（Zeng 等人，2024）和 TIS-DPO（Liu 等人，2025）。表 13 所示的实验结果表明，尽管采用片段级采样策略的 DPO 和 PPO 训练相较于 SFT模型取得了中等程度的性能提升，但与采用细粒度对齐损失的 ALPO 相比，仍存在显著差距。ALPO 的表现也显著优于这两种词元级方法。这验证了分段采样策略和细粒度对齐损失对于局部偏好优化任务的有效性。

表 13：替代解决方案的实验评估结果。C 表示粗粒度，F 表示细粒度。第 1 名和第 2 名的结果分别以 **蓝色** 和 **橙色**表示。

模型	训练	采样	en⇒zh 准确性	en⇒zh 自然度	en⇒zh 生动性	zh⇒en 准确性	zh⇒en 自然度	zh⇒en 生动性
黄金参考	-	-	*83.6*	*82.9*	*71.6*	*82.9*	*80.2*	*73.2*
GPT-4o	-	-	*89.5*	*82.4*	*59.5*	*88.6*	*82.9*	*64.6*
DeepSeek-R1	-	-	*90.4*	*85.6*	*70.7*	*88.5*	*85.6*	*73.6*
Qwen2.5-14B	SFT	-	86.5	82.1	59.2	85.2	80.1	54.9
	DPO	C	87.1	82.2	63.3	86.1	81.1	60.2
		F	88.2	82.5	69.4	**86.3**	83.1	68.4
	PPO	F	88.3	**83.2**	**70.1**	85.9	82.3	**70.1**
	TDPO	F	87.4	82.7	64.0	86.0	81.3	61.2
	TIS-DPO	F	**88.5**	83.1	65.9	86.2	**83.3**	62.4
	ALPO	F	**90.6**	**84.2**	**76.6**	**88.3**	**86.8**	**81.6**

C.5 ALPO 在多轮交互中的进一步应用

C.5.1 概述

为验证 ALPO 方法的普适性，我们在另一项局部偏好优化任务上进行了实验：社交语言智能体多轮交互（Zhou 等，2024；Wang 等，2024a；

Lu等人，2024；Kong等人，2025）。该任务要求大语言模型智能体在交互环境中，通过多轮交互并根据角色档案、情境场景和私有社交目标，动态调整生成策略。其目标是使智能体能够在特定社交场景（例如谈判、协作、竞争）中更有效地实现预定义目标（例如说服、达成一致），同时维持或改善对话双方之间的关系。

C.5.2 方法

对于多轮交互任务，我们主要通过两个步骤完成偏好对齐数据采样：1）错误定位；2）偏好数据采样。对于训练集中的多轮对话，我们首先使用 GPT-4o 识别错误轮次的起始位置，遵循与 Kong 等人（Kong等人，2025）相同的标准：1）该轮对实现角色目标至关重要；2）目标完成度与角色之间的关系仍有提升空间。

如果两个标准均满足，我们便将该轮之前的对话历史作为初始采样前缀 p_1，然后使用我们的大语言模型智能体为每一轮依次采样 5 个单轮回复。在第 i 次采样迭代中，我们使用 GPT-4o 根据采样得到的响应判断当前轮是否至关重要：主要包含客套话或类似内容的响应被视为不重要，其指示函数 $\mathbf{1}(p_i)$ 设为 0；否则设为 1。随后，我们选择目标与关系综合得分最高的响应作为偏好回复 y_i^w，得分最低的作为非偏好回复 y_i^l，并优先考虑目标完成度而非关系。第 i 次迭代的前缀 p_i 变为 p_{i-1}，y_{i-1}^w（包括对话者行为，此处略去）。最后，我们采用以下简化的 ALPO 损失函数进行对齐训练：

$$\mathcal{L}_{\text{alpo}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_{1:n}^w,y_{1:n}^l)\sim\mathbb{D}}\left[\sum_{i=1}^{n}\mathbf{1}(p_i)\cdot\log\sigma\left(\beta\log\frac{\pi_\theta(y_i^w\mid p_i)}{\pi_{\text{ref}}(y_i^w\mid p_i)}-\beta\log\frac{\pi_\theta(y_i^l\mid p_i)}{\pi_{\text{ref}}(y_i^l\mid p_i)}\right)\right], \tag{13}$$

where $\mathbb{D}$ 表示数据集，π_θ 和 π_{ref} 均从行为克隆（BC）模型初始化 s.

C.5.3 实验

我们从 SOTOPIA-π（Wang 等，2024a）的 410 个场景中选取 100 个用于行为克隆（每个场景 10 个角色对），并选取 310 个场景用于对齐（每个场景 8 个角色对）。我们采用 SOTOPIA（Zhou 等，2024）作为测试集，其中包含 90 个场景，每个场景有 5 个角色对，共计 450 个自对话任务和 900 个非自对话任务。我们在相同的基线系统上评估目标与关系维度，实验结果见表 14。基线系统的结果引自 Kong 等人（Kong等人，2025）。

表 14：各种方法在 SOTOPIA 的目标与关系维度上的表现。第 1 名和第 2 名的结果分别以**蓝色**和**橙色**表示。

方法	Self-Chat		GPT-4o	
	Goal	Rel	Goal	Rel
GPT-3.5-turbo	6.38	1.36	7.19	2.05
GPT-4o-mini	6.98	2.11	7.44	2.36
GPT-4-turbo	8.18	2.96	7.92	2.79
GPT-4o	7.90	2.67	7.90	2.67
LLaMA-8B	7.24	1.94	7.70	2.49
LLaMA-8B+行为克隆	7.81	3.05	7.53	2.78
LLaMA-8B+行为克隆+偏好监督微调	7.76	3.05	7.65	2.88
LLaMA-8B+行为克隆+DPO (Rafailov 等，2024年)	7.95	3.28	7.80	2.97
LLaMA-8B+行为克隆+ETO (Song 等人，2024)	8.29	3.39	8.02	3.03
LLaMA-8B+行为克隆+DMPO（Shi 等人，2024）	8.28	3.37	8.00	2.98
LLaMA-8B+行为克隆+SDPO（Kong 等人，2025）	**8.56**	**3.69**	**8.13**	**3.16**
LLaMA-8B+行为克隆+ALPO	**8.41**	**3.56**	**8.21**	**3.19**

实验结果证明，ALPO 在其他应用任务中的有效性。ALPO 采用从局部最优解逐步过渡到全局最优解的渐进式优化策略，同时利用

指示函数自适应地确定每个片段是否参与损失计算。该方法能够有效处理多片段局部优化任务。

D 学说：自适应局部偏好优化

本节将形式化定义局部偏好优化问题，并验证 ALPO 的有效性以及通用偏好优化方法的局限性。需要注意的是，为了便于阐述，本节使用的符号可能与前文有所不同。

D.1 局部偏好优化问题

对于给定的输入 $x \in \mathcal{X}$（$\mathcal{X}$表示输入空间），输入语言模型 π_θ，假设其由 n个相互关联的片段组成，即 $x = (x_1, \cdots, x_n)$。相应地，其输出 $y \in \mathcal{Y}$（$\mathcal{Y}$表示输出空间）也由 n个相互关联的片段组成，即 $y = (y_1, \cdots, y_n)$，其中 x_i与 y_i一一对应。此外，y_i的生成会受到 $y_1, \cdots, y_{i-1}$的影响（请注意，这种影响不仅源于语言模型的自回归特性，也可能源于输出片段之间的语义依赖等因素），即：

$$\pi_\theta(y \mid x) = \prod_{i=1}^{n} \pi_\theta(y_i \mid x, y_1, \cdots, y_{i-1}). \tag{14}$$

考虑一个片段级偏好指标 $r(x_i, y_i)$（奖励信号或其他定量度量），用于评估片段输出 y_i 与其对应片段输入 x_i的优化目标的契合程度。局部偏好优化问题的目标是调整参数 θ，使 π_θ得到优化，从而最大化 $r(x_i, y_i)$。该目标的形式化表达如下：

$$\theta^* = \arg\max_\theta \mathbb{E}_{x \sim p(x)} \Big[\mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \big[\sum_{i=1}^{n} r(x_i, y_i) \big] \Big], \tag{15}$$

其中，$p(x)$表示输入 x的真实分布。

D.2 结果监督偏好优化方法的无效性

我们以 DPO（Rafailov 等，2024年）为例，说明通用偏好优化方法在解决局部偏好优化问题时的局限性。在执行局部偏好优化时，DPO 首先根据 $r(x_i, y_i)$，为 $x = (x_1, \cdots, x_n)$标注偏好响应 $y^w = (y_1^w, \cdots, y_n^w)$ 和非偏好响应 $y^l = (y_1^l, \cdots, y_n^l)$。这些响应本质上满足：

$$\pi_\theta(y^w \mid x) = \prod_{i=1}^{n} \pi_\theta(y_i^w \mid x, y_1^w, \cdots, y_{i-1}^w), \pi_\theta(y^l \mid x) = \prod_{i=1}^{n} \pi_\theta(y_i^l \mid x, y_1^l, \cdots, y_{i-1}^l). \tag{16}$$

然后，DPO 通过单阶段策略优化方法优化 π_θ：

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y^w, y^l) \sim \mathbb{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y^w \mid x)}{\pi_{\text{ref}}(y^w \mid x)} - \beta \log \frac{\pi_\theta(y^l \mid x)}{\pi_{\text{ref}}(y^l \mid x)} \right) \right], \tag{17}$$

其中，π_{ref} 表示冻结的参考模型。对于损失函数中的对比项，计算：

$$\log \frac{\pi_\theta(y^w \mid x)}{\pi_{\text{ref}}(y^w \mid x)} - \log \frac{\pi_\theta(y^l \mid x)}{\pi_{\text{ref}}(y^l \mid x)} = \sum_{i=1}^{n} \left[\log \frac{\pi_\theta(y_i^w \mid x, y_{1:i-1}^w)}{\pi_{\text{ref}}(y_i^w \mid x, y_{1:i-1}^w)} - \log \frac{\pi_\theta(y_i^l \mid x, y_{1:i-1}^l)}{\pi_{\text{ref}}(y_i^l \mid x, y_{1:i-1}^l)} \right], \tag{18}$$

其中，$y_{1:i-1}^w = (y_1^w, \cdots, y_{i-1}^w)$ 和 $y_{1:i-1}^l = (y_1^l, \cdots, y_{i-1}^l)$。这种方法的局限性如下：

- **对齐前缀上缺乏细粒度对比**：第i个片段的生成概率取决于其自身的偏好前缀或非偏好前缀（$y_{1:\,i-1}^w$ 或$y_{1:\,i-1}^l$），而非共享前缀。如果 y_i^w与 y_i^l 的前缀存在显著分歧（例如风格或主题分支不同），则第 i个片段处的对比不再是相同条件下的公平比较。

- **无法按片段施加训练信号**：DPO对完整序列（y^w，y^l）进行整体偏好判断，使模型无法获得片段级纠正指导。具体而言，尽管模型能够识别 y^w 相对于 y^l的整体优势，但它缺乏局部反馈来优化各个片段上的决策。
- **梯度稀释与噪声**：当 y^w 与 y^l 之间的分歧仅出现在特定片段时，全局对比会汇总许多无关片段（生成内容几乎相同）产生的梯度。这会淹没关键的对齐信号，妨碍有效识别并纠正未对齐片段。

总之，在局部偏好优化任务中，DPO 表现不佳，这是由前缀发散导致的生成条件未对齐，以及缺乏逐步学习"正确与错误"决策所致（其他结果监督偏好优化损失也存在这一局限）。

D.3 自适应局部偏好优化

与原始 DPO 不同，ALPO 使用 $y^w_{1:i-1}$ 作为前缀，为每个片段标注偏好回复 y^w_i和非偏好回复 y^l_i，而不是生成完整的响应序列。训练期间，第 i 个片段的前缀固定为 $p_i = (x, \hat{}y_1, \cdots, \hat{}y_{i-1})$，其中 $\hat{}y_i$通过预定式前缀混合策略获得。这确保了每个片段的比较都基于相同的偏好前缀进行，从而消除了"偏好前缀与非偏好前缀"之间的不公平竞争。在相同前缀下，偏好对齐损失会比较以 p_i 为条件的 $\pi_\theta(y^w_i \mid p_i)$ 和$\pi_\theta(y^l_i \mid p_i)$。最后，对分段损失求和，从而对每个片段施加偏好约束：

$$\mathcal{L}_{\text{alpo}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y^w_{1:n}, y^l_{1:n})\sim\mathbb{D}}\left[\sum_{i=1}^{n} w(p_i)\cdot \log\sigma\left(\beta_i \log\frac{\pi_\theta(y^w_i \mid p_i)}{\pi_{\text{ref}}(y^w_i \mid p_i)} - \beta_i \log\frac{\pi_\theta(y^l_i \mid p_i)}{\pi_{\text{ref}}(y^l_i \mid p_i)}\right)\right], \tag{19}$$

其中 $w(p_i)$ 和 β_i 是自适应超参数，根据特定标准决定每个片段对优化过程的贡献。

使用计划式前缀混合来获得 $\hat{}y_{1:i-1}$ 作为前缀的原因在于，随着优化的进行，模型倾向于生成 y^w_i，从而使训练后的分布与 $y^w_{1:i-1}$ 路径更加一致。由于 $y^w_{1:i-1}$ 也是最可能被遵循的前缀路径，这确保模型在其最频繁访问的前缀下学会生成偏好片段 y^w_i 而非 y^l_i。这与强化学习和偏好优化中的一个常见原则相一致：训练期间的状态（或前缀）分布应与最终策略最可能访问的分布相匹配（同策略）（Christiano等人, 2017；Sutton, 2018；Ouyang 等, 2022年）。形式上，在训练过程中我们旨在最小化：

$$\mathbb{E}_{p\sim d_{\theta^*}}\Big[\mathcal{L}_{\text{alpo}}\big(\pi_\theta(\cdot \mid p)\big)\Big], \tag{20}$$

其中 d_{θ^*} 表示由最终模型诱导的前缀分布。由于 d_{θ^*} 在收敛时很可能包含偏好前缀，选择 $\hat{}y_{1:i-1}$ 作为前缀实质上是从偏好路径中采样（或直接使用）状态，从而确保与最终分布 d_{θ^*} 的一致性。

E 讨论

在本节中，我们将进一步讨论 ALPO 方法。

E.1 结论

视觉媒体行业作为人类文化交流与传播的重要媒介，正见证着自动化和工业化成为未来的关键趋势。在此背景下，我们提出了 ALPO——一种新颖的翻译模型训练范式，采用了 LLM-as-a-Judge 和偏好对齐等技术。通过全面的实验和理论分析，我们验证了 ALPO 的有效性。我们相信，这项研究不仅将为视觉媒体技术的进步作出积极贡献，也将推动其他领域对领域特定翻译模型的研究。

E.2 建议

我们想为使用 ALPO 的技术人员提供一些开发建议：

- 使用大语言模型进行偏好标注时，应在多轮推理中打乱译文的顺序，并对结果取平均，以获得稳健的评估结果。
- 训练提示应使用最适合所选骨干网络的语言，例如，通义千问系列模型使用中文提示，LLaMA 系列模型使用英语提示。
- 我们的开源实现揭示了更多处理细节，包括术语翻译、字幕分段和语义完整性保留。具体实现请参阅源代码。

字幕文本本质上是多模态内容，与媒体节目的视觉和听觉元素密切交织。视频提供了有关角色、场景和情节发展的补充信息，而音频则传递情感和语气线索。因此，未来的研究应着重利用这些多模态数据资源来改进字幕翻译，例如，借助视频和音频模态为字幕行添加说话人标注，并利用视听上下文辅助字幕翻译。

E.3 局限性

ALPO 存在以下局限性：

- 字幕翻译所需的视觉媒体数据及相关版权主要集中在大型企业手中，造成了数据可获取性障碍。
- 与其他最先进的大语言模型类似，ALPO 缺乏在翻译过程中整合多模态信息（例如视频）的能力。

F 提示与指令

本节介绍大语言模型和人工评估者所使用的输入和输出格式。

F.1 翻译大语言模型的输入与输出

下面展示 zh⇒en 翻译模型的提示与响应格式（其他语言也采用类似格式）。红色文本仅用于大语言模型上下文学习。节目原始字幕和翻译字幕按文本框所示格式排版，用于监督微调和 ALPO 模型训练，以及其他基线大语言模型上下文学习。

翻译模型的 zh⇒en 翻译提示示例。

<指令>
请根据以下要求，将以下多行中文影视台词翻译成英语：1. 英语翻译应通俗易懂、口语化，且语言风格与中文台词一致。2. 英语翻译必须富有表现力、生动传神，有效传达中文原文的氛围、情感和语气。3. 输出翻译时请保留行号，并确保译文行数与中文原文行数一致，不要合并任何行。

<示例>
原文：1.请给你我一点和平相处
的时间 2.我们必须谈谈
. . .

34.秘术在你体内涌动
35.请你离海克斯飞门远点

翻译结果：1.请给我们一点和平相处的时间。
2.我们必须谈谈。……34.奥术在你体内涌动。
35.离Hexgates远点。

<任务>
请按照要求完成以下翻译。
原文：
1.我听说了你的病情
2.我真的很遗憾，孩子
. . .
34.令人心痛的是，越是光芒万丈的人
35.生命能量消耗得就越快

翻译结果：

SFT模型的zh⇒en翻译响应演示。

1.我听说了你的预后。2.我非常难过，我的男孩。……34.令人悲伤的事实是 35.最耀眼的人往往燃烧得最快。

F.2 评估大语言模型的输入

我们在下方给出用于大语言模型评估 zh⇒en 翻译生动性的提示格式（其他语言采用类似格式）。使用大语言模型评估某一行的生动性时，我们会提供其周围的上下文行。

评估大语言模型的 zh⇒en 提示示例。

<指令>
请为一条中文字幕行的多条英文译文赋予生动性分数（0-100的整数）。原始中文行及其周围上下文将作为参考提供，分别以 [待评估] 和 [上下文] 标注。翻译生动性的评分原则如下：原则1（准确性）：译文可以适度意译，但必须在传达原意时保持合理的准确性，不得增删（权重：30%）；原则2（口语自然度）：评估译文是否使用适合字幕语境的自然口语，并有效传达原有的情感、氛围和语气（权重：30%）；原则3（表现力）：评估译文是否具有情感共鸣、运用生动措辞，并具备能够吸引观众的文学品质（权重：40%）。

评分标准（生动性）：100：翻译极富表现力、感染力，情感层次丰富，能深深打动观众。50：翻译表现力中等，情感表达较为含蓄，仅部分传达了预期感受。0：翻译不准确，缺乏情感深度或表现力，

无法引起观众共鸣。

<示例>
中文原文：[上下文] 被你到处践踏的遗迹 可是无价之宝[上下文] 是
无法用价值来衡量的珍宝[待评估] 历史虽然会重演 但是人类是无法
回到过去的[上下文] 看来你是不会明白的[上下文] 明...明白...我明
白了

英语翻译：
译文 A：历史会重演，但我们无法回到过去。

译文 B：历史常常回响，然而我们却无法
让时光倒流。

...

评估分数：
{"A": 70, "B": 92, "C": 85, ...}

<任务>
根据评判标准，完成以下评估。
中文原文：
[上下文] 我的贡献注定是转瞬即逝 连您都会很快遗忘
[上下文] 我见过的学生很多
[待评估] 令人心痛的是 越是光芒万丈的人 越会更快地燃尽生命的能量
[上下文] 我不知道你还是个艺术家
[上下文] 我的事你不知道的多着呢

英语翻译：
译文 A：令人悲伤的是，最耀眼的人
往往燃烧得最快。

译文 B：令人心碎的事实是，那些最耀眼的人往往会更快耗尽生命的能量。

...

请注意，输出评分时需采用 JSON 格式：{"A": score, "B": score, "C": score, ...} 评估分数
（仅输出评分，不要输出其他内容）：

F.3 质量评估说明与提示

要对翻译质量进行人工评估，必须向评估器提供明确的指令，说明评估视角、标准和格式。这些指令会直接影响评估器在质量评估过程中的关注重点。提供给评估器的指令见第一个文本框，而用于借助大语言模型进行多维自动评估的提示则列于后续三个文本框中。

翻译质量人工评估说明。

[评估标准]

1. 准确性
评估视听字幕翻译的准确性时，请考虑以下几个方面：

- **语义等效性**：评估原字幕的含义是否在译文版本中得到准确传达，以及源字幕的语义内容是否被精准

表达于目标字幕中。

·**语法正确性**：评估翻译字幕的语法准确性，包括句子结构、时态、语态及其他语法方面。

·**术语翻译**：评估专有名词是否得到准确翻译，并确保保留原术语的语义和上下文。

2. 自然度
评估视听字幕翻译的自然度时，请考虑以下维度：

·**连贯性**：评估翻译字幕读起来是否像由目标语言母语者撰写，并评估句子之间的逻辑关系。

·**可读性**：评估翻译字幕是否易于阅读和理解，以及用词和表达是否符合目标语言的惯例。

·**流畅性**：评估翻译字幕是否通顺流畅、句子结构是否合理，以及是否存在明显的语法错误或不自然表达。

3. 生动性
评估视听字幕翻译的生动性时，请考虑以下维度：

·**风格一致性**：评估翻译字幕是否保留了原文的风格和特点，包括角色语气和情感细微差别是否一致。

·**表现力**：评估翻译是否传达了原台词的精髓和氛围，避免机械直译，从而让观众更容易理解并产生兴趣。

·**情感**：评估译文是否忠实传达角色的情感，是否符合场景和角色的上下文，并能在情感上引起目标语受众的共鸣。

[任务]

对于每组原字幕，均提供了两种不同的译文（A 和 B）。请参考 [评估标准] 中规定的多个评估维度，对每组原字幕的两种译文进行评估。请通过勾选 [A 更好]，[、]，B 更好 [或] A 和 B 之间无显著差异]，分别标明译文 A 和 B 的评估结果。请注意，您只需评估每组字幕的整体表现，而不是字幕中的每一行。

英译中字幕翻译准确性评估的提示。

[英译中字幕翻译准确性评估]

请使用 0 到 100 的整数分数，评估以下英译中字幕翻译的准确性。翻译准确性评分标准包括评估中文翻译是否准确传达了英语字幕的原意。此外，请注意英语原字幕中的术语（例如人名、地名、组织名称、物品名称等）是否在中文翻译中得到了准确翻译。

评分标准（准确性）：100：翻译完全准确，涵盖所有信息，表达连贯，且专有名词翻译准确。50：翻译基本准确，仅有轻微遗漏

或上下文不清，且专有名词翻译部分有误，但整体含义仍然可以理解。0：翻译严重失真，误解原文的主要含义，且专有名词翻译质量较差。

英文原句：
这样的罪行不能被忽视。这个男孩必须受到惩罚。违反 Ethos 就应遭到驱逐，但我能理解一个年轻人改变世界的梦想。

. . .
是啊，我必须承认，他的学说很引人入胜。如果危险的念头不曾激
发想象，我们就永远不会误入歧途。

中文翻译：这样严重的罪行不能轻易放过 违
反社会共识的人确实应该遭到驱逐 但我也能
体会一个年轻人梦想改变世界的雄心
. . .
我不得不说，他那套学说非常有意思。如果
危险的念头不曾引人遐想，也就没有误入歧
途一说了。

注意：请以 JSON 格式输出：{"Score": 评估分数} 分数（仅输出分数，无需进一步解释）：

英语⇒中文字幕翻译自然度评估提示。

[英译中文字幕翻译自然度评估]

请使用 0 到 100 的整数分数，对以下英译中文字幕翻译的自然度进行评分。评估字幕翻译的自然度时，应考虑译文是否充分考虑了文化背景和上下文等因素，以及语言表达是否自然流畅，符合中文语法结构和用词习惯。译文应易于理解，并贴近中国观众的文化

e

和表达习惯。

评分标准（自然度）：
100：翻译充分考虑了上下文和文化背景，语言流畅自然，符合中文表达习惯，且
没有语法或用词错误。

50：翻译考虑到了上下文，整体基本流畅，但有些表达可能略显生硬或不自然，也可能存在轻微的语法错误。0：翻译未能有效考虑上下文或文化因素，语言不流畅，存在许多语法错误，表达生硬且难以理解。

英文原句：
这样的罪行不能被忽视。这个男孩必须受到惩罚。违背 Ethos 应被驱逐，但我能理解一个年轻人改变世界的梦想。

. . .
是啊，我必须承认，他的学说很吸引人。如果危险的念头无法激发
想象，我们就永远不会误入歧途。

中文翻译：如此严重的罪行不能轻易放过。
违反社会共识的人确实应该遭到驱逐，但我
也能体会一个年轻人梦想改变世界的雄心。

. . .
我不得不说，他那套学说非常有意思。如果
危险的念头不曾引人遐想，也就没有误入歧
途一说了。

注意：请以 JSON 格式输出：{"Score": 评价分数} 分数（只需输出分数，无需进一步解释）：

英⇒中字幕翻译生动性评估提示。

[英译中文字幕翻译生动性评估]

请为下面的英译中文字幕翻译生动性打分，使用 0 到 100 的整数。翻译生动性评分不考虑翻译的准确性；只评估译文是否富有表现力、情感丰富，以及是否更能吸引观众。

评分 g 标准（生动性）：
100：翻译极具表现力，感染力十足，情感丰富，能够深深打动观众。50：翻译有一定的表现力，情感表达较为平淡，但仍能传达一些情感。0：翻译缺乏表现力和情感，无法引起观众的任何情感共鸣。

英文原句：
这样的罪行不可忽视。这个男孩必须受到惩罚。违反 Ethos 就应被驱逐，但我能体谅一个年轻人想要改变世界的梦想。

. . .
是啊，我得承认，他的学说很吸引人。如果危险的念头不能激发想
象力，我们就永远不会误入歧途。

中文翻译：这样严重的罪行不能轻易放过。
违反社会共识的人确实应该遭到驱逐，但我
也能体会一个年轻人梦想改变世界的雄心。

. . .
我不得不说，他那套学说非常有意思。要是
危险的念头不曾引人遐想，也就无所谓误入
歧途了。

注意：你需要以 JSON 格式输出：{"Score": 评价分数} 分数（仅输出分数，无需进一步解释）：