arXiv:2603.01691v1 [cs.CL] 2 Mar 2026

预印本

为低资源语言构建强大的指令语言模型

Domen Vreš、Tjaša Arčon、Timotej Petrič、Dario Vajda、Marko Robnik-Šikonja 和 Iztok Lebar Bajec

斯洛文尼亚卢布尔雅那卢布尔雅那大学计算机与信息科学学院 *通讯作者。电子邮箱: domen.vres@fri.uni-lj.si

摘要

大语言模型（LLM）已成为自然语言处理以及人工智能领域的重要工具。目前的开源模型主要以英语文本训练，因此在低资源语言和文化方面的性能较差。我们介绍了将大语言模型成功适配至低资源语言所需的一系列方法学方法，并以斯洛文尼亚语为例加以展示。我们推出 GaMS3-12B，这是一款拥有 120 亿参数的斯洛文尼亚语生成模型，并证明它是在其参数规模范围内性能最佳的斯洛文尼亚语开源模型。我们通过对 Gemma 3 模型进行三阶段持续预训练，再进行两阶段监督微调（SFT），使模型适配斯洛文尼亚语。模型使用共计 1400 亿个斯洛文尼亚语、英语、波斯尼亚语、塞尔维亚语和克罗地亚语预训练词元，以及超过 20 万个英语和斯洛文尼亚语监督微调样例进行训练。我们在 Slovenian-LLM-Eval 数据集、英语到斯洛文尼亚语翻译任务以及斯洛文尼亚语大语言模型竞技场上评估 GaMS3-12B。结果表明，所述模型在这三种场景下均优于 12B 的 Gemma 3，并且在斯洛文尼亚语大语言模型竞技场上的表现可与规模大得多的商业模型 GPT-4o 相媲美，胜率超过 60%。

关键词： GaMS，大语言模型，语言适配，低资源语言

1. 引言

大语言模型（LLM），尤其是 GPT 模型等生成式大语言模型，彻底改变了自然语言处理（NLP），推动了我们对人类语言的理解，并提升了生成语言的能力。随着这一领域迅速发展，Gemma（Team，2024 年）、Llama（Dubey 等人，2024 年）、Nemotron（英伟达，2025年）等新型开源解码器型Transformer大语言模型不断推出。这些模型主要使用高资源语言（尤其是英语）进行训练，而对斯洛文尼亚语等低资源语言的注意力和性能则较为有限。为改善这一状况，我们介绍了 **GaMS3-12B**（**斯**洛文**尼**亚语**生**成**模**型），这是一个面向斯洛文尼亚语、拥有 120 亿参数的开源生成模型。我们使用一套斯洛文尼亚语基准测试 Slovenian-LLM-Eval（CJVT，2024年）、英语到斯洛文尼亚语翻译以及斯洛文尼亚语大语言模型竞技场[1]对 GaMS3-12B 进行评估。结果表明，该模型明显优于基础版 Gemma 3 模型，充分说明可以提升大语言模型在许多其他低资源语言上的表现。鉴于此类开发工作正在进行或计划开展，我们对开发过程的介绍具有超越单一语言的显著泛化价值。

为低资源语言训练大语言模型的主要挑战在于数据稀缺。例如，Llama 4 模型的训练数据包含 30 万亿个词元，而目前可用的真实斯洛文尼亚语语料库仅包含约 400 亿个词元，规模约为其千分之一

1. https://arena.cjvt.si/en/leaderboard

。这意味着，从头开始为斯洛文尼亚语（以及其他低资源语言）训练大语言模型并不可行。因此，我们采用预训练的 Gemma 3 模型，并针对斯洛文尼亚语进行适配。我们进行了三个持续预训练（CPT）阶段，随后进行了两个监督微调（SFT）阶段。为增加训练数据量，我们还纳入了克罗地亚语、波斯尼亚语和塞尔维亚语等相近语言的文本，这有助于提升模型性能。此外，我们通过对高质量斯洛文尼亚语 PDF 文件进行光学字符识别来生成更多数据，例如来自斯洛文尼亚顶尖大学知识库的文档，以及国家图书馆数字档案中的文档。我们还翻译了维基百科和 Nemotron 数据集等高质量英语语料库。

我们的贡献如下：

- GaMS：一个最先进的斯洛文尼亚语开源生成模型；
- 一套用于斯洛文尼亚语文档光学字符识别的开源代码，可迁移至其他低资源语言；
- 清晰阐述了数据生成与准备方法，以及模型训练步骤，使我们的工作能够迁移至其他低资源语言。

开发低资源语言的开源模型，对于保护这些语言的自主性、文化认同和主权至关重要。目前，这些语言只有在商业模型中才能获得充分支持，因此想要使用这些语言的用户不得不将数据发送给大型公司。相比之下，对于拥有开放式大语言模型的语言，公司和公共机构可以在本地部署开源模型，从而完全掌控客户的数据驻留。此外，斯洛文尼亚语 GaMS 模型的创建过程开放且文档齐全，也为其他低资源语言提供了有益经验。

本文其余部分安排如下。第2节介绍与大语言模型语言适配相关的研究。第3节详细说明我们的持续预训练方法和数据。第4节介绍监督微调数据和方法。第5节介绍训练环境、硬件资源和超参数。第6节评估我们的模型，第7节总结全文并提出未来工作的设想。

2. 相关工作

大多数最先进的语言模型（Brown 等人，2020 年；Comanici 等人，2025 年；Dubey 等人，2024 年；Team，2024 年）都采用相同的流程进行训练。模型首先经历预训练阶段，在海量语料库上接受训练，以预测下一个词元。随后进入监督微调（SFT）阶段，训练模型响应用户的提示。最后，使用基于人类反馈的强化学习（RLHF）等方法，训练模型生成符合用户偏好（例如更安全、事实更准确且更有帮助）的输出。我们采用了类似的流程，但由于缺少斯洛文尼亚语的偏好数据，因此以对话增强型监督微调取代了强化学习。

多语言模型旨在将大语言模型的优势扩展到训练数据不足的语言。这些模型通常以英语为基础，并加入其他语言的训练数据。斯洛文尼亚语表现最好的开源多语言模型包括 EuroLLM（Martins 等，2025年）、Gemma 2（Team，2024 年）、Gemma 3（Team，2025年）、DeepSeek（Guo 等，2025年）、Qwen（Yang 等，2025年）、Llama 4（Meta，2025年）、Mistral（Mistral-AI 等，2025年）和 Apertus（Apertus 等，2025年）。然而，DeepSeek、Llama 4 以及 Mistral 和 Qwen 的最大版本通常难以本地部署，因为它们的参数数量庞大，需要大量 GPU 显存。另一方面，Mistral 和 Qwen 的较小版本在斯洛文尼亚语上的表现明显较差，不及其他规模相近的模型。对于斯洛文尼亚语而言，上述模型系列中最值得关注的是 Gemma，因为其性能出色；而 EuroLLM 则因专注于欧盟语言而值得关注。EuroLLM 从头开始，使用由以下部分组成的

数据混合进行训练：50% 的英语文本、5% 的代码/数学数据以及 45% 的其他语言数据。这些其他语言涵盖了欧洲的大多数语言。斯洛文尼亚语数据占其他语言数据的 1%。Gemma 模型预训练所用的斯洛文尼亚语数据比例尚不明确，但其占比足够高，因此 Gemma 3 的 27B 版本在 Slovenian-LLM-Eval（CJVT，2024年）基准测试套件中的表现优于其他多语言模型（见表8）。这证明了我们选择 Gemma 3 作为起始模型是合理的。然而，Gemma 3 模型仍能从更多斯洛文尼亚语数据的训练中获益，这一点从 GaMS3 相较于 Gemma 3 12B 的性能提升中清楚可见。

过去几年，一些解码器型英语模型被改编用于特定语言。最早的此类模型包括面向瑞典语的 GPT-SW3（AI-Sweden 2024）、面向中文的中文 LLaMA（Cui、Yang 和 Yao 2023）和 Open-Chinese-LLaMA（OpenLMLab 2023），以及面向葡萄牙语的 Gervasio（Santos 等 2024）。然而，这些模型要么是从头开始训练的（GPT-SW3），要么在未进行嵌入迁移的情况下扩展词汇，导致知识损失（中文 LLaMA 和 Open-Chinese-LLaMA），要么针对目标语言进行了指令微调（监督微调）（Gervasio）。相比之下，我们首先对 Gemma 3 模型进行持续预训练；我们认为，这对于使模型适应斯洛文尼亚文化至关重要，随后再进行了监督微调（SFT）。近期，英伟达启动了主权人工智能倡议，支持欧洲、中东和亚洲的主权语言模型开发者。除 GaMS 系列外，该倡议最值得关注的代表还包括 Falcon（Zuo 等 2025年）和 Bielik（Ociepa 等 2025年）。Falcon 从头开始训练，侧重于多语言能力；而 Bielik 团队采用了与我们类似的方法。他们以预训练的 Mistral 模型为基础，通过持续预训练、监督微调和强化学习将其适配为波兰语模型。尽管 Bielik 在斯洛文尼亚语上的表现尚可，但仍未达到斯洛文尼亚语专用模型的水平。

已有多个面向斯洛文尼亚语开发的大语言模型。编码器型和编码器—解码器型模型包括 CroSloEngual BERT（M. Ulčar 和 M. Robnik-Šikonja，2020年）、SloBERTa（Ulčar 和 Robnik Šikonja，2021年）以及 SloT5（Matej Ulčar 和 Marko Robnik-Šikonja，2023）。最早适配斯洛文尼亚语的解码器型模型包括 SlovenianGPT（Gordić，2024年）、GaMS-1B（Vreš 等，2024年）和 TinySLlama（zID4si，2024年）。由于架构较旧、上下文较短且训练数据有限，与当前模型相比，这些模型在实际应用中的价值较低。不过，它们为斯洛文尼亚语专业化提供了宝贵见解，展现了将模型与斯洛文尼亚语相近的语言（如克罗地亚语）结合训练（CroSloEngual BERT）的优势，以及词汇修改和嵌入迁移（GaMS-1B）的影响。目前，最具代表性的斯洛文尼亚语专用解码器型大语言模型是基于 Gemma 2 的 GaMS 模型（Vreš 等，2025），参数规模为 9B 和 27B；此外还有基于 Mistral NeMo 的 Zlatorog 12B（Savodnik，2025b）。其中表现最佳的是 GaMS-27B-Instruct-Nemotron（Petrič、Vreš、Bajec、Arčon 等，2025a）。与 GaMS3 相比，这些模型的训练数据都少得多（约 15B 个词元），上下文也短得多（Zlatorog 的上下文窗口最长，为 16k 个词元），因此无法处理较长文本。另一方面，GaMS3 使用约 140B 个词元进行训练，并支持 131k 个词元的上下文窗口。

3. 持续预训练

将大语言模型适配到特定语言可分为两个阶段：持续预训练和监督微调。本节介绍第一阶段，第二阶段将在第4节中介绍。

持续预训练（CPT）阶段对于提升模型对目标语言和文化的理解至关重要。我们此前对 GaMS 模型的研究（Vreš等，2024年；Vreš 等，2025年）表明，经过持续预训练后，模型在斯洛文尼亚语基准测试中的表现有所提升。对于 GaMS3，我们将持续预训练分为三个阶段：平行对齐、基础持续预训练和长文本持续预训练。每个阶段采用不同的数据组合，旨在以特定方式增强基础模型对

斯洛文尼亚语和文化的理解。我们在第 3.1 节介绍平行对齐阶段，在第 3.2 节介绍基础持续预训练阶段，在第 3.3 节介绍长程持续预训练阶段。为了获取更多持续预训练数据，我们对高质量的斯洛文尼亚语 PDF 文件进行了光学字符识别。我们在第 3.4 节介绍 OCR 流程。为了改进训练过程并提高效率，我们采用了多种数据准备技术，并在第 3.5 节对其进行介绍。

3.1 并行对齐

平行对齐阶段旨在明确对齐模型内部的英语和斯洛文尼亚语。此阶段的灵感来自 C. Li 等人（2024年），他们在预训练阶段之后结合多语言对比学习和跨语言指令微调来实现语言对齐。前者旨在提高模型对源语言和目标语言中同一句子的表示之间的相似度。后者则向模型提供源语言的提示词，并要求模型以目标语言作答。我们将平行对齐提前至第一个持续预训练阶段，以弥合 Gemma 3 的训练分布（以英语为主）与我们的训练分布（以斯洛文尼亚语为主）之间的差距。在训练上一代 GaMS 模型（GaMS-9B 和 GaMS-27B）时，这种方法已展现出良好的效果。

为了进行平行对齐，我们结合使用了以下英语和斯洛文尼亚语平行语料库：

- **人工翻译语料库**：KAS 摘要（Žagar 等，2022年a）、DGT（Tiedemann，2012）和 MaCoCu（Bañón 等，2023）；
- **机器翻译语料库**：维基百科

在此，我们将采用机器翻译方式把 Markdown 格式的英文维基百科（Savodnik（2025a））翻译成斯洛文尼亚语，从而将维基百科转换为并行语料库；翻译所用模型为 GaMS-9B-SFT-DPO-Translator（Vajda、Vreš 和 Robnik-Šikonja，2025年）。并行语料库的分词后数据规模如表 1 所示。在此阶段，我们使用 65k 词元的上下文窗口训练模型。

表 1. 并行对齐语料库中的词元数量和训练文档数。数字是在使用 Gemma3 分词器进行分词、序列打包和填充后显示的。

语料库	词元数量	文档数量	总百分比
DGT	804,847,616	12,281	6.3 %
MaCoCu	430,374,912	6,567	3.4 %
KAS	31,391,744	479	0.2 %
维基百科	11,529,093,120	175,920	90.1%
总计	**12,795,707,392**	**195,247**	

我们在三个层级上对齐了英语和斯洛文尼亚语文本。在段落级对齐中，平行文档的段落被拼接在一起。一种语言的一个段落后面紧接着另一种语言中的同一段落，然后再接第一种语言的下一段落，依此类推。在文档级对齐中，两种不同语言的文档被拼接成单个训练样例。在独立文档中，不同语言的文档被用作不同的训练样本。对齐层级是根据数据集格式和平均文档长度选择的。

3.2 基础持续预训练

基础持续预训练阶段背后的理念是，首先使用较短的上下文执行持续预训练，以提高训练效率。在此阶段，我们使用了 65k 个词元的上下文窗口。我们选择这样的

上下文窗口，是为了使训练仍能在可用的图形处理器硬件（即 LEONARDO Booster 分区）上完成。在这一阶段，我们选择不使用更短的上下文窗口，因为我们的预训练组合中包含较长的文档。此外，Gemma 3 支持 131k 个词元的上下文窗口，我们不希望因使用过短的序列长度而削弱其长上下文能力。在这一阶段，模型使用斯洛文尼亚语、英语、克罗地亚语、波斯尼亚语和塞尔维亚语的通用语料库进行训练。我们纳入英语语料库以防止遗忘，同时加入克罗地亚语、波斯尼亚语和塞尔维亚语语料库，是因为这些语言与斯洛文尼亚语具有文化和语言上的相似性，并且能够为代码混合文本提供额外支持。先前的研究（M. Ulčar 和 M. Robnik-Šikonja 2020年；Vreš等，2024年）表明，加入此类语言有助于提升模型在斯洛文尼亚语上的性能，同时也能使模型在这些相近语言上取得较好的性能。

我们在基础持续预训练的数据混合中结合了以下类型的数据：

- **英伟达 Nemotron 预训练数据集**（英伟达等，2025年；Mahabadi等，2025年）：英伟达发布的英语预训练数据集组合。我们从已发布的数据集中选取了部分拆分，并从各个拆分中随机抽取数据，以确保持续预训练数据的大部分为斯洛文尼亚语。
- **来自先前项目的斯洛文尼亚语语料库集合**：metaFida（Erjavec，2023）。
- **学术数据**：KAS（Žagar等，2022b）、斯洛文尼亚各大学知识库中的文档（主要为理学学士、文学学士、理学硕士、文学硕士和博士论文），截止年份为2018年。
- **新闻数据**：Trendi（Kosem等（2025）），此阶段的数据截止日期为2023年12月（长程持续预训练阶段使用了更新的数据）。• **法律数据**：整合来自 PISRS 等公共来源的斯洛文尼亚语法律数据 [2] 以及《官方公报》 [3]。
- **医疗数据**：斯洛文尼亚语医疗书籍、博客、期刊、论文和文章的汇编。
- **国家图书馆数据**：来自斯洛文尼亚国家与大学图书馆（NUK）数字档案的数据。所用数据不受版权保护，由旧报纸、期刊和书籍组成。其中也有一些现代科学期刊或论文。
- **维基百科**：斯洛文尼亚语、克罗地亚语、波斯尼亚语和塞尔维亚语维基百科的 Markdown 格式转储。所有语言的转储日期均为2025年1月1日。
- **网页抓取数据**：CLASSLA-web（Ljubešić和Kuzman，2024年）和FineWeb2（Penedo等，2025年）中的斯洛文尼亚语部分。我们使用基于ReWire 方法的数据生成提示词和Gemma 3 27B，对部分CLASSLA语料库（100万份文档）进行了增强（Nguyen等，2025年）；
- **网页 PDF 文件**：FinePDFs 语料库中斯洛文尼亚语、克罗地亚语、波斯尼亚语和塞尔维亚语的数据子集（Kydlíček、Penedo和Werra，2025年）。此阶段使用了约75%的数据，其余25%留待长程持续预训练阶段使用。

基础持续预训练语料库的分词后数据规模见表 2。

3.3 长文本持续预训练

在最后的长文本持续预训练阶段，上下文窗口增加到 131k 词元，以覆盖完整的 Gemma 3 上下文窗口。在此阶段，仅使用了以下最高质量的语料库：

- **Nemotron 语料库**：与基础持续预训练阶段相比，我们省略了 Math-3 和预训练代码拆分。对于其余拆分，我们使用了比基础持续预训练阶段更小且互不重叠的子集。此外，我们使用 GaMS-9B-SFT-DPO-Translator 将 Math-4 和预训练监督微调拆分翻译成斯洛文尼亚语。翻译前者是为了增加斯洛文尼亚语数学数据的数量，翻译后者则是为了添加一些斯洛文尼亚语预训练监督微调数据，这可能会带来益处。

2. https://pisrs.si 3. https://www.uradni-list.si

表 2. 基础持续预训练语料库的词元数量和训练文档数量。数字为使用 Gemma 3 分词器进行分词、序列打包和填充后的结果。Wikipedia-Yugo 指斯洛文尼亚语、波斯尼亚语、克罗地亚语和塞尔维亚语维基百科语料的合称。NUK 指斯洛文尼亚国家与大学图书馆。

语料库	语言	词元数量	文档数量	总百分比
Nemotron-Pretraining-Code	英语	1,952,120,832	29,787	1.9 %
Nemotron-CC-Math-v1-4Plus	英语	2,526,937,088	38,558	2.5 %
Nemotron-CC-Math-v1-3	英语	1,210,908,672	18,477	1.2 %
Nemotron-Pretraining-SFT-v1	英语	3,718,316,032	56,737	3.7 %
Nemotron-CC-v2-HighQuality-Synthetic	英语	10,479,403,008	159,903	10.4 %
Nemotron-CC-v2-DiverseQA	英语	8,631,353,344	131,704	8.6%
FinePDFs-Bos	波斯尼亚语	4,815,912,960	73,485	4.8%
FinePDFs-Hrv	克罗地亚语	9,541,124,096	145,586	9.5%
FinePDFs-Srp	塞尔维亚语	8,119,844,864	123,899	8.0%
FinePDFs-Slv	斯洛文尼亚语	5,925,044,224	90,409	5.9%
Trendi	斯洛文尼亚语	1,737,687,040	26,515	1.7%
CLASSLA-Web-2024	斯洛文尼亚语	4,256,432,128	64,948	4.2 %
Sl-Legal	斯洛文尼亚语	1,697,710,080	25,905	1.7%
Sl-Med	斯洛文尼亚语	1,598,095,360	24,385	1.6%
metaFida	斯洛文尼亚语	4,591,910,912	70,067	4.6 %
FineWeb2	斯洛文尼亚语	13,890,289,664	211,949	13.8 %
KAS	斯洛文尼亚语	2,726,035,456	41,596	2.7 %
NUK	斯洛文尼亚语	12,784,041,984	195,069	12.7 %
Wikipedia-Yugo	斯洛文尼亚语、克罗地亚语、波斯尼亚语、塞尔维亚语	673,775,616	10,281	0.7 %
总计		**100,876,943,360**	**1,539,260**	

• **FinePDFs**：我们使用了剩余的 25% 斯洛文尼亚语、克罗地亚语、波斯尼亚语和塞尔维亚语 FinePDFs 数据。
• **新闻数据**：我们使用了 2024年1月至2025年7月期间的 Trendi 语料库中的新闻文章。

• **学术数据**：由于 KAS语料库中的学术作品仅涵盖截至2018年发表的内容，我们决定对从斯洛文尼亚各大学的知识库中收集的 PDF 文件进行光学字符识别。我们纳入了这些作品 P发表于2019年至2024年间。我们将此语料库称为 P我们的 KAS-Extension。
• **斯洛文尼亚语数学数据**：我们对两份斯洛文尼亚数学与物理期刊 Presek [4] 和 OMF [5]进行了光学字符识别。由于这两份期刊合计不足 1 亿词，我们使用 Nemotron Math 的译文构建了斯洛文尼亚语数学语料库。

长文本持续预训练语料库的分词后数据规模如表3所示。

3.4 斯洛文尼亚语 *PDF* 文件的光学字符识别

我们对来自三个不同来源的 PDF 文件进行了光学字符识别：

• **国家图书馆数据**：我们从国家图书馆获得了大量 PDF 格式的旧报纸、书籍、期刊、文章及其他材料，这些材料大多是原作的扫描件。部分获得的 PDF 文档还附带了 txt 格式的光学字符识别版本（使用 ABBYY 工具创建）。由于这些材料年代久远，整体质量较低、格式不佳且存在大量伪影（书籍尤其如此），我们决定使用现代光学字符识别工具重新进行光学字符识别。在我们的所有数据集中，这一数据集规模最大，也是最具挑战性的。

• **斯洛文尼亚大学的毕业论文**：我们从排名前 3 的斯洛文尼亚大学知识库中收集了 2019 年至 2024 年期间发布的 PDF 文件。尽管这个数据集同样规模庞大，但它

4. http://www.presek.si/predstavitev.php
5. http://www.obzornik.si

表3. 长篇持续预训练语料库的词元数量和训练文档数量。数字为使用 Gemma 3 分词器进行分词、序列打包和填充后的结果。

语料库	语言	词元数量	文档数量	总百分比
Nemotron-CC-Math-v1-4Plus	英语	1,087,373,312	8,296	5.4 %
Nemotron-Pretraining-SFT-v1	英语	1,231,945,728	9,399	6.1 %
Nemotron-CC-v2-HighQuality-Synthetic	英语	2,634,285,056	20,098	13.1 %
Nemotron-CC-v2-DiverseQA	英语	1,237,975,040	9,445	6.2 %
FinePDFs-Bos	波斯尼亚语	1,614,282,752	12,316	8.0 %
FinePDFs-Hrv	克罗地亚语	2,385,248,256	18,198	11.9 %
FinePDFs-Srp	塞尔维亚语	2,074,345,472	15,826	10.3 %
FinePDFs-Slv	斯洛文尼亚语	1,969,618,944	15,027	9.8 %
Trendi	斯洛文尼亚语	610,533,376	4,658	3.0 %
KAS-Extension	斯洛文尼亚语	2,256,404,480	17,215	11.2 %
Sl-Math	斯洛文尼亚语	1,456,078,848	11,109	7.2%
Nemotron-预训练-SFT-翻译	斯洛文尼亚语	1,553,858,560	11,855	7.7 %
总计		**20,111,949,824**	**1,539,260**	

比国家图书馆的数据更不具挑战性，因为这些 PDF 文件是以数字方式生成的，且其内容具有更加结构化的版式。

- **斯洛文尼亚数学与物理期刊**：我们收集了两份斯洛文尼亚数学与物理期刊的 PDF 文件：Presek 和 Obzornik za matematiko in fiziko。它们的版式比大学学位论文稍具挑战性，但这些 PDF 文件均为数字化生成，且数据集规模较小。

我们采用了三阶段光学字符识别流程，并将在开源许可证 [6] 下发布源代码。第一阶段，使用光学字符识别模型将 PDF 文件转换为 Markdown。第二阶段，使用单独的语言模型对 Markdown 进行后处理。第三阶段，使用 NeMo Curator 工具清理 Markdown。我们在图 1 中以流程图展示了该流程，并在接下来的三个小节中进行说明。

3.4.1 光学字符识别阶段

我们初步测试了不同的光学字符识别模型，并通过人工检查和内部基准测试，选出了三个适用于我们任务的模型：

- **Marker** [7] 是一个 PDF 转 Markdown 转换库，可将整份文档转换为 Markdown（因此无需逐页进行后处理，因为 marker 库已处理了这一步）。除了最终的 Markdown 外，该库还会提取图表和 PNG 格式的图像，并将它们纳入 Markdown 中。该库在内部运行多个不同的视觉模型，并可通过商业 API 或本地服务器进行大语言模型后处理。我们在此未启用大语言模型选项，因为测试表明它并未显著改善结果，却显著拖慢了转换速度。我们使用的是该库的 `1.6.2` 版本。在所用的三种光学字符识别方法中，这种方法运行速度最快，但对不同版式和复杂格式最为敏感。

- **Llama-4-Maverick** （Meta 2025年）是 Meta 发布的多模态多语言大语言模型，具有出色的光学字符识别性能。我们将文档逐页发送给该模型，将 PDF 文件转换为 Markdown。每一页都以 Base64 编码图像的形式发送，并使用附录 3.1 中的提示词指示模型以 Markdown 格式提取文本。Llama

6. https://github.com/GaMS-Team/local_ocr
7. https://github.com/datalab-to/marker

4 在学术作品上的性能优于另外两个模型，这类作品的版面结构定义明确。此外，它还能准确地从 PDF 文件中提取方程和表格。然而，在处理复杂版面时（尤其是文字分布杂乱的旧报纸），它的表现不如 Nanonets。它对计算资源的要求也最高，运行速度最慢。它需要至少 640 GB 的总显存（例如，一台配备 8 块各有 80 GB 显存的 H100 GPU 的完整节点，或四块各有 180 GB 显存的 B200 GPU）。我们在内部 GPU 集群的一台 DGX B200 节点上部署了它。

• **Nanonets-OCR-s** （Nanonets 2025）是 Qwen2.5-VL-3B-Instruct 多模态模型的微调版本。该模型在一个由研究论文、金融文件、法律文件、医疗文件、税务表格、收据和发票组成的精选数据集上进行了微调。它支持方程转录（LaTeX）、图像描述、水印提取以及复杂表格的识别。与 Llama 4 类似，我们将每份 PDF文档逐页发送给模型，并编码为 Base64 图像。我们使用了模型卡片中建议的默认提示词（确切提示词见附录 3.2）。与 Llama 4 相比，该模型运行速度更快，计算需求更低。它能很好地处理困难的版面，但由于模型规模的原因，往往会在文本中插入伪影，并可能开始循环。幸运的是，此类伪影和循环很容易检测，并且可以在过滤阶段予以去除。

需要注意的是，我们于2025年5月启动了光学字符识别流程，当时许多现有的 OCR 模型（如 DeepSeekOCR（Wei、Sun 和 Li 2025年））尚未发布。此外，GPT 和 Gemini 等商业模型的表现可能更好，但由于数据规模和隐私问题，使用这些模型并不可行。

我们使用 Llama 4 对学术作品以及数学与物理学期刊进行光学字符识别。对于国家图书馆的数据，我们采用了以下方法。对于已标注报纸/期刊/书籍标题的作品（我们没有从国家图书馆获得任何元数据），我们首先使用 marker 库进行光学字符识别。随后，我们检查每份报纸/期刊/书籍的若干示例；如果质量较差，则根据版式复杂程度，改用 Llama 4（较简单的版式）或 Nanonets（最具挑战性的版式）重复处理。对于没有任何元数据的 PDF 文件，我们直接通过 Nanonets 进行处理，因为在这三个模型中，它的质量与速度比最高。

3.4.2 后处理阶段

由于 PDF 文件是按单页分别送入 Llama 4 和 Nanonets 处理的，因此在后处理阶段，需要使用 Gemma 3 27B it 模型将这些页面重新合并为一个文档。我们选择 Gemma 3，是因为它体量较小，能够在合理的时间内处理大量数据；与其他同等规模的开源模型（GaMS 除外）相比，它更擅长理解斯洛文尼亚语，而且上下文窗口比前几代 GaMS 模型更大。后处理分为两个步骤。第一步，让 Gemma 3 将每一页分类为正文或模板内容。这一步很有必要，因为每本书开头都有许多空白页或模板内容页。此外，文档中可能包含仅有图像或广告的页面，也可能有整页在光学字符识别过程中被转换为单个伪影的情况。这些页面需要在合并前移除。

分类完成后，将成对的页面合并为一个文档。具体做法是遍历各页，将当前页的最后一段和下一页的第一段发送给 Gemma 3。随后，让 Gemma 3 根据以下情形之一合并页面：

• 当前页面的最后一段是页脚。在这种情况下，合并后的版本会删除这一段。

• 下一页的第一段是页眉。在这种情况下，合并后的版本会删除该段。

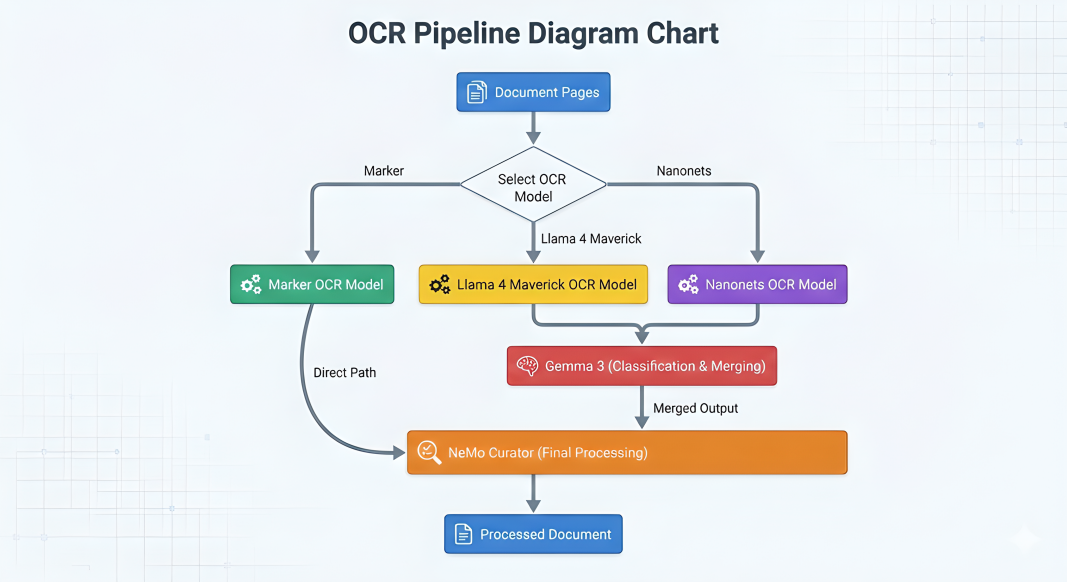

图 1。我们的三阶段光学字符识别流程。每份 PDF文档首先通过 OCR 模型转换为 Markdown。对于 Llama 4 和 Nanonets，文档会逐页进行光学字符识别，然后使用 Gemma 3 27B it 对页面进行后处理。使用 marker 库时，这两个阶段会在库内合并。最后阶段使用 NeMo Curator 筛选器进行文本清洗。该图表由 AI 使用 Nano Banana Pro 工具生成。

- 当前页的最后一段以一个带连字符的单词结尾，该单词会在下一页继续。在这种情况下，合并时会将两个单词直接拼接起来，不保留“-”符号。
- 当前页的最后一个段落与下一页的第一个段落属于同一段落。在这种情况下，它们会合并为一个段落。

- 当前页的最后一个段落与下一页的第一个段落属于不同段落。在这种情况下，页面会通过换行合并。

3.4.3 过滤阶段

在最后阶段，Markdown 文档会经过一系列 NeMo Curator [8] 修饰器。所有文档都会经过以下筛选器：

- **图像移除器**会移除嵌入 Markdown 中的图像。
- **换行规范化器**会将连续的换行字符截断为最多两个。
- **Unicode重格式化器**会修正编码不正确的 Unicode 字符。
- **ČŠŽ校正器**：会将斯洛文尼亚语变音字符 č、š 和 ž 的错误形式（例如ˇc）替换为正确形式。

对于 Nanonets，我们增加了两项额外的过滤规则，以去除模型产生的伪影。第一项规则会删除过长的段落（超过 15,000 个字符的段落），因为这表明其中可能存在伪影，或重复的词语或句子（模型开始循环生成）。第二项规则会删除重复的段落。如果某个段落在一份文档中重复出现超过 100 次，则说明模型可能陷入循环，只保留首次出现的段落。

8. https://github.com/NVIDIA-NeMo/Curator

3.5 最终数据准备

在训练之前，我们进行了额外的数据处理，以提高训练的质量和效率。通过这一步数据处理，我们将每个训练样本的长度设定为与各持续预训练阶段使用的序列长度完全一致。使用 NeMo 框架 9 时，这一步至关重要，否则框架本身可能会将训练文档拆分到多个训练样本中。也就是说，如果一份文档在达到序列长度之前就结束，框架会接上下一份文档的开头，从而随机拆分该文档。除了上下文不足之外，这还会导致输入序列出现问题，因为下一个训练实例不会以大多数大语言模型所要求的序列起始（BOS）标记开头。对于 Gemma 类型模型，这尤其成问题，因为不包含 BOS 标记会大幅降低其性能。为避免这种情况，我们采用了以下流程。

首先，我们对每份文档进行分词，并将长度超过目标序列长度的文档拆分为更小的单元。不同语料库中的单元有所不同，可以是句子、段落或章节。随后，我们将连续且总长度不超过上下文窗口长度的单元合并为子文档。接着，采用基于启发式方法的序列打包，将拆分后的文档和子文档合并为单个训练样例，且其长度不超过上下文窗口长度。最后，所有短于上下文窗口长度的训练样本都会填充序列结束（EOS）标记。

4. 监督微调

持续预训练之后，大语言模型适配特定语言的下一阶段是监督微调（SFT）阶段，该阶段能够提升模型在各类任务上的最终性能。CPT 阶段构建模型的知识，并能显著增强其语言能力和对文化的理解；而 SFT 阶段则使模型能够表达其知识。GaMS 模型的上一代充分展现了 SFT 阶段的影响：GaMS-9B-Instruct（Vreš 等，2025）与 GaMS-9B-Instruct-Nemotron（Petrič、Vreš、Bajec、Arčon 等，2025b）之间的差异仅在于 SFT 数据，但它们在斯洛文尼亚语大语言模型竞技场上的排名却相差显著。GaMS-9B-Instruct 排名靠后，而 GaMS-9B-Instruct-Nemotron 则名列前茅，甚至击败了商业模型 GPT-4o 和 Gemini-2.0-Flash。

对于 GaMS3，我们将指令微调分为两个监督微调阶段：通用指令微调和聊天微调。通用指令微调阶段基于 GaMS-Instruct 数据集，聊天微调阶段则基于 GaMS-Nemotron-Chat 数据集。这两个数据集都是专为斯洛文尼亚语模型的指令微调而设计的。我们在第4.1节介绍 GaMS-Instruct 数据集的生成过程，并在第4.2节介绍 GaMS-Nemotron-Chat 数据集的准备过程。

通用指令微调阶段旨在让模型学习多种不同类型的任务和主题。经过这一阶段后，模型应能胜任下游任务。我们在这一阶段使用了以下数据集组合：

- **GaMS-Instruct**：我们将新创建的通用 GaMS-Instruct 数据集（由 ClosedQA、OpenQA 和写作任务组成）与基于数字人文的 GaMS-Instruct-DH 1.0（Šorn 等人，2024）数据集结合使用。

- **GaMS-Lex** 是一个关于斯洛文尼亚语词典编纂的问答集合，内容来源多样，包括基于 ssj500k（Krek 等人，2021）的题目、来自 Jezikovna svetovalnica 10 的题目，以及提示词马拉松期间手动编写的题目。

- **Nemotron-Post-Training-v2**（Nathawani、Ding 等，2025年）：我们使用了该数据集中的科学、技术、工程和数学子集和数学子集，因为这些子集涵盖了我们的混合数据集中其他数据集覆盖不足的主题。我们

9. https://github.com/NVIDIA-NeMo/NeMo
10. https://svetovalnica.zrc-sazu.si

使用 Gemini-2.5-Flash 将部分示例翻译成斯洛文尼亚语。我们同时纳入了英语和斯洛文尼亚语示例；不过，我们确保英语示例与翻译成斯洛文尼亚语的示例不同。

- **斯洛文尼亚代码反馈**（Machidon 和 Vreš 2025）是基于英语 CodeFeedback 及类似数据集翻译而成的代码数据集。该数据集使用 Gemini- 2.5-Flash 翻译为斯洛文尼亚语，包含不同的任务类型和编程语言。代码数据集通常被认为有助于大语言模型训练，因为其中包含结构化数据以及各种算法的许多实现。

各数据集中的样本数量见表4。对于大多数数据集，我们采用了 90/10 的训练验证集划分比例。例外情况是 Nemotron 数据集中的斯洛文尼亚语部分，我们对数据进行了下采样，以防止模型对其特定任务类型过拟合；另一个例外是 GaMS-Lex，由于数据量有限，我们仅将其用于训练。

表4. 通用指令微调阶段各数据集的训练样本和验证样本数量。

数据集	语言	训练样本数量	验证样本数量
GaMS-Lex	斯洛文尼亚语	1,884	0
GaMS-Instruct-ClosedQA	斯洛文尼亚语	10,825	1,202
GaMS-Instruct-OpenQA	斯洛文尼亚语	28,704	3,189
GaMS-Instruct-Writing	斯洛文尼亚语	9,056	1,006
GaMS-Instruct-DH 1.0	斯洛文尼亚语	9,135	1,015
Nemotron-SFT-v2-Math-En	英语	9,000	1,000
Nemotron-SFT-v2-Math-Sl	斯洛文尼亚语	10,000	1,000
Nemotron-SFT-v2-STEM-En	英语	9,000	1,000
Nemotron-SFT-v2-STEM-Sl	斯洛文尼亚语	10,000	1,000
SlCode	斯洛文尼亚	10,000	1,000
总计		**107,604**	**11,412**

在对话微调阶段，模型被训练为更好地作为通用聊天模型。GaMS-Nemotron-Chat 数据集在上一代 GaMS 模型上已在这一方面展现出良好效果。该数据集引导模型变得更加用户友好，使其更有帮助、阐述更详尽，并能更好地遵循所要求的回复格式。为使模型更安全，我们在此阶段加入了一小组安全提示。这些安全提示由人工撰写，我们计划未来逐步扩充该数据集。对话微调阶段的数据集规模见表5。我们对 GaMS-Nemotron-Chat 数据集采用了 90/10 的训练-验证划分，而由于数据集规模较小，所有安全样本均被纳入训练集。

表5. 对话微调阶段各数据集的训练样本与验证样本数量。

数据集	语言	训练样本数量	验证样本数量
GaMS-Safety	斯洛文尼亚语	459	0
GaMS-Nemotron-Chat	斯洛文尼亚语、英语	88,126	9,791
总计		**88,585**	**9,791**

4.1 GaMS-Instruct 数据集

由于斯洛文尼亚语等低资源语言缺乏用于大语言模型指令微调的大规模高质量数据集，我们创建了 GaMS-Instruct，这是一个面向斯洛文尼亚语的综合性指令微调数据集，

并遵循最先进的构建指南，强调构建涵盖多样化任务和主题的均衡数据集的重要性。该数据集包含 53,983 个指令-响应对，旨在支持斯洛文尼亚语大语言模型的微调。

4.1.1 数据集构建

我们将数据集划分为三个主要部分，分别对应用户提交的常见提示类型。指令—响应对的类型分布和规模基于近期研究结果，见表4。数据集包含以下任务类型：

- **开放式问题**，涵盖广泛的主题和子主题；
- **封闭式问题**，根据斯洛文尼亚语源文本生成；
- **写作任务**，针对预先定义的文本类型。

开放式问答 初步实验表明，不受限制地生成开放式问题会导致主题分布不均。为解决这一问题，我们手动选择主题，用于生成开放式问题任务。为确保跨领域覆盖广泛且均衡，主题选择依据IPTC媒体主题分类法 11、维基百科的分类体系 12 以及《大英百科全书》的在线分类 [13]。我们总共生成了20个涵盖广泛领域的大类。每个类别进一步细分为30个更具体的子主题。在指令生成过程中，每个子主题都会被重复使用多次，以确保每个主题都有足够数量的示例。

封闭式问答 封闭式问题任务基于三类现有的斯洛文尼亚语文本来源生成，答案直接从源文本中提取：

- **Trendi**（Kosem等（2025）），这是一个包含斯洛文尼亚语新闻和网络内容的语料库，共有3,000个示例；
- **KAS**（Žagar等（2022b）），这是一个包含斯洛文尼亚语学术论文的语料库，共有3,000个示例；
- **斯洛文尼亚语维基百科**（Savodnik（2025a）），包含 6,028 个示例。

写作任务 我们设计了写作任务的分布，使其能够反映大语言模型用户对不同文本类型的需求频率。我们根据可能的写作任务的频率，将其分为三类：

- **高频任务**（例如电子邮件、论文和故事），每种生成 400 次；
- **中频任务**（例如手册、食谱和演讲稿），每种生成 140 次；
- **写作任务**（例如贺卡、菜单或请愿书），每种生成 25 次。

除了生成特定类型的文本外，我们还制作了涵盖改写、编辑、简化等写作相关活动的任务。

4.1.2 指令生成

为了针对所述各个领域生成提示，我们改编了用于多语言指令生成的 Self-Instruct框架（Wang等，2023年），并特别聚焦于斯洛文尼亚语。与依赖一组固定的人工编写英语种子任务的原始方法不同，我们的方法直接根据特定主题输入或源文档生成指令，而不使用预定义的种子示例。为提升数据集的指令多样性，大约一半的

11. https://iptc.org/standards/media-topics 12. https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories 13. https://www.britannica.com/

数据集由 Gemini 2.0 Flash 生成，另一半则由 GPT-4o 生成。结构化提示系统会动态创建系统提示和用户提示，并采用了从 NeMo Curator 工具包中选取的提示模板。为确保新颖性并减少冗余，我们使用 ROUGE-L 相似度，将新生成的指令与此前生成的指令进行比对并筛选。对于原本不包含答案的任务（例如写作任务），我们还会提示模型生成有效的响应，从而得到完整的指令-响应对，可用于指令微调。

4.2 GaMS-Nemotron-Chat

为提升模型的对话质量，我们构建了 GaMS-Nemotron-Chat 数据集（Petrič、Vreš、Bajec 与 Robnik-Šikonja，2025年）。我们通过翻译 Nemotron-Post-Training-Dataset-v1（Nathawani、Gitman 等人，2025年）中的英语样本来创建该数据集。该数据集包含针对 LMSYS Chat 1M（Zheng 等人，2024年）中真实世界用户提示的回复。数据集中的回复由 Qwen3-235B-A22B（Yang 等人，2025年）模型重新生成，该模型在 ArenaHard-v2（T. Li 等人，2025年）等基准测试上取得了良好的性能。为确保多样性并减少冗余，我们从包含约 273,000 个样本的 LMSYS Chat 1M Clean 子集（OpenLeecher，2024年）出发。随后，我们在每个类别（例如编程、创意写作）内应用 MinHash LSH 筛选（Indyk 与 Motwani，1998年），采用 0.65 的相似度阈值和 256 次置换，最终筛选出去重后约 80,000 个互不相同的高质量样本。

我们使用 GaMS-27B-Instruct 模型将去重后的示例翻译成斯洛文尼亚语，并采用附录 1.1 所示的自定义提示词模板，以保留对话结构（即 <用户> 和 <助手> 标签）。为确保合成数据的质量，我们实施了长度比过滤器。我们仅保留译文长度与源文本长度之比处于 [0.73 到 1.35] 范围内的译文。事实证明，这种启发式方法能够有效避免幻觉和翻译不完整的问题；在该范围之外，这些问题时有发生。在初期阶段，我们发现了两个关键问题：由于源数据的影响，模型经常自称为“Qwen”；此外，用英语提示时，它会用斯洛文尼亚语作答。为解决身份问题，我们找出了 694 段提及“Qwen”的对话，并使用附录 1.2 所示的提示词重新生成回复。这种方法明确将模型的身份设定为 GaMS。

为缓解语言混用问题并保留模型以英语作答的能力，我们在 80,000 条斯洛文尼亚语译文的基础上，加入了约 20,000 条来自 Nemotron-Post-Training-Dataset-v1 的英文原始示例，并确保这些示例与译文子集没有重叠。因此，最终的 GaMS-Nemotron-Chat 数据集约包含 100,000 条示例，其中斯洛文尼亚语与英语的比例为 80/20。

5. 训练设置

本节介绍将大语言模型适配到资源较少的斯洛文尼亚语的技术细节。我们以 Gemma 3 12B（预训练版本）为起点。我们使用 NeMo 2.0 框架进行持续预训练（CPT）（使用官方 NVIDIA NeMo 容器的 25.07 版本 [14]）。选择 NeMo 是因为它支持 Gemma 3，并提供多种优化和并行化方案。结合官方容器，它能够在多个 GPU 和节点上实现高效扩展。进行监督微调（SFT）时，我们改用 Transformers（版本 `4.57.3`）和 DeepSpeed [15]（版本 `0.18.2`），因为我们发现使用 NeMo 对 Gemma 3 进行 SFT 时存在一些质量问题。由于这一阶段的计算强度较低，因此可以在优化程度较低的环境中完成。

在持续预训练阶段，我们使用了以下优化技术和并行策略：

14 =25 07 .
15 https://catalog.ngc.nvidia.com/orgs/nvidia/containers/nemo?version . .
https://huggingface.co/docs/transformers/deepspeed

- **张量并行（TP）**将每个线性层和注意力头拆分到多个 GPU 上。
- **序列并行**将丢弃和层归一化等层拆分到各个张量并行秩上。
- **激活重计算**（也称为激活检查点）：反向传播时，不存储所有激活值，而只存储每个注意力层之前的激活值，并在反向传播过程中重新计算层内的激活值。这可以降低显存用量。

对于监督微调，我们使用了 DeepSpeed ZeRO 阶段 2。所有阶段的超参数如表 6 所示。所有训练阶段均使用`bf16`精度。

表 6. 持续预训练和监督微调阶段的训练超参数。在基础持续预训练阶段，我们采用逐步增加批次大小的策略：961 步后，批次大小从 128 增至 192；又经过 600 步后，批次大小再次增至 256。两个监督微调阶段均训练了 3 轮；不过，根据验证损失选择了第二轮训练后的检查点。两个监督微调阶段的超参数均通过网格搜索确定。

	CPT			SFT	
超参数	并行对齐	基础持续预训练	长文本持续预训练	指令微调	聊天微调
模型并行	TP 8	TP 8	TP 8	DeepSpeed ZeRO 阶段 2	DeepSpeed ZeRO 阶段 2
数据并行	64	64	16	8	8
批次大小	128	128 → 192 → 256	64	64	64
微批次大小	1	1	1	8	8
学习率调度器	带预热的余弦	带预热的余弦	带预热的常数	带预热的余弦	带预热的余弦
最小学习率	5×10^{-7}	5×10^{-7}	/	1×10^{-6}	1×10^{-6}
最大学习率	5×10^{-6}	5×10^{-6}	5×10^{-6}	5×10^{-6}	5×10^{-6}
预热步数	150	1000	500	1000	1000
恒定步数	200	1000	/	0	0
轮次	1	1	1	2	2

我们使用以下硬件进行训练：

- **LEONARDO Booster** 是一台 EuroHPC 超级计算机，其 Booster 分区拥有 3200 个配备 4 张 A100 64GB GPU 的节点。单个节点内的 GPU 通过 600 GB/s 的 NVLink 连接，节点之间则通过 2x200 Gb/s 的 InfiniBand 连接。我们成功将训练扩展到 128 个节点（集群设置的硬性上限为 256 个节点）。我们使用 LEONARDO Booster 进行平行对齐阶段的训练，以及基础持续预训练阶段的大部分训练。在 LEONARDO 上，我们为训练耗用了约 14 万 GPU 小时。
- **内部集群 FRIDA**：我们使用了属于内部 FRIDA 集群的一台配备 8 张 B200 180GB GPU 的 DGX 节点，以及一台配备 8 张 H100 80GB GPU 的 DGX 节点。当 LEONARDO 上的资源耗尽时，我们使用这两个节点完成了基础持续预训练阶段的一小部分训练。训练共耗用了约 120 个 B200 GPU 小时和 960 个 H100 GPU 小时。
- **NVIDIA DGX CloudLepton** 是一个统一的 AI 平台，可通过全球云服务提供商网络将开发者与数万块 GPU 相连接。我们获准使用 16 个 8×H200 141GB节点。我们使用 Lepton 完成基础持续预训练阶段及所有后续阶段。我们获配了 4 万 GPU 小时。

6. 评估

我们通过三种场景对 GaMS3 进行了评测。在第 6.1 节中，我们介绍模型在 Slovenian-LLM-Eval 基准测试套件上的评测结果，该套件用于测试模型的通用知识。在第 6.2 节中，我们评测一项特定的生成任务，即模型在英语到斯洛文尼亚语翻译方面的性能。最后，在第 6.3 节中，我们使用 Slovene-LLM-Arena 评测模型的生成能力。

6.1 Slovenian-LLM-Eval

Slovenian-LLM-Eval（CJVT 2024年）是一个基准测试集合，属于 LM评测工具集（Gao 等，2024年）的一部分。英语基准测试的斯洛文尼亚语版本由机器翻译而成。其中一部分由 Aleksa Gordić 及其社群使用谷歌翻译服务翻译，随后再用 GPT-4o 进行润色。另一部分则作为欧洲大语言模型排行榜（Thellmann 等，2024年）的一部分，使用 DeepL 翻译。我们使用了以下基准测试：ARC-Easy、ARC-Challenge、BoolQ、GSM8K、HellaSwag、OpenBookQA、PIQA、TruthfulQA 和 Winogrande。此外，我们还请斯洛文尼亚语母语者人工检查并修正 ARC-Easy、ARC-Challenge、OpenBookQA 和 Winogrande。

我们将 GaMS3 12B 与规模相近的多语言开源模型进行了比较。所有模型均使用指令微调版本。各项基准测试的结果见表 7。与 Zlatorog 和 GaMS-9B 相比，GaMS3 在各项基准测试中表现良好。其表现略逊于规模是它两倍的 GaMS-27B。与 Gemma 3 12B 相比，GaMS3 在几乎所有基准测试中的表现都显著更好。仅在 GSM8K 和 TruthfulQA 上例外。我们推测，这些任务表现较差的原因是采用了少样本评估。这表明有必要提升 GaMS3 处理长输入的性能。我们还可以扩充监督微调数据集，将少样本任务也纳入其中。另一个有趣的发现是，在由斯洛文尼亚语母语者人工校正的基准测试中，斯洛文尼亚语专用模型的表现几乎胜过所有开源模型。这进一步凸显了对其他基准测试进行人工校正的必要性。

表 7. Slovenian-LLM-Eval 中斯洛文尼亚语专用模型（上）和多语言开源模型（下）在各项基准测试上的结果。分数越高越好。每项基准测试中表现最佳的模型分数以粗体标示，第二佳模型的分数加下划线。Acc. 表示准确率，EM 表示严格精确匹配。n-shot 列显示添加到提示词中的已解答示例数量。由斯洛文尼亚语母语者手动校正的基准测试以 † 标记。

基准测试	指标	n-shot	GaMS3-12B	GaMS-9B-Nemotron	GaMS-27B-Nemotron	Zlatorog-12B
ARC 挑战 †	Acc.	0	0.5265	0.5171	**0.5444**	0.4676
ARC 简单 †	Acc.	0	0.7744	0.7546	**0.7875**	0.7079
BoolQ	Acc.	0	0.8523	0.8471	0.8618	0.8321
GSM8K	EM	5	0.6892	0.6277	0.6892	0.5921
HellaSwag	Acc.	0	0.5111	0.5140	**0.5428**	0.5373
OpenBookQA †	Acc.	0	0.3940	**0.4080**	0.3980	0.4060
PIQA	Acc.	0	0.7149	0.7062	**0.7252**	0.7247
TruthfulQA MC1	Acc.	6	0.3807	0.3415	0.3672	0.3647
TruthfulQA MC2	Acc.	6	0.5396	0.5307	0.5314	0.5383
Winogrande †	Acc.	0	0.7056	0.7190	**0.7206**	0.7017

基准测试	指标	n-shot	Gemma 3 12B	Gemma 3 27B	EuroLLM-22B	Qwen3-30B-A3B	Apertus-8B	Bielik-11B-v3.0
ARC 挑战 †	Acc.	0	0.4514	0.5137	0.5026	0.4386	0.4556	0.4923
ARC 简单 †	Acc.	0	0.6936	0.7475	0.7433	0.6427	0.7071	0.7231
BoolQ	Acc.	0	0.8526	**0.8630**	0.8278	0.8590	0.8330	0.8596
GSM8K	EM	5	0.7430	**0.8006**	0.5148	0.7202	0.4375	0.7339
HellaSwag	Acc.	0	0.4728	0.5235	0.4931	0.4256	0.4888	0.5212
OpenBookQA †	Acc.	0	0.3520	0.3700	0.3880	0.2860	0.3420	0.3780
PIQA	Acc.	0	0.6616	0.7106	0.6904	0.6246	0.6893	0.6937
TruthfulQA MC1	Acc.	6	0.3978	**0.4076**	0.3537	0.3953	0.3917	0.3794
TruthfulQA MC2	Acc.	6	0.5827	**0.5881**	0.5229	0.5845	0.5574	0.5517
Winogrande †	Acc.	0	0.6559	0.6717	0.6551	0.6204	0.6496	0.6922

为便于比　为比较各模型在所有数据集上的表现，我们针对每项基准指标将模型从1（最佳）到

10（最差）进行排名，并计算模型在所有基准测试中的平均排名。结果见表8。GaMS3的排名仅次于规模大一倍的GaMS-27B-Nemotron和Gemma 3 27B。它显著优于Gemma 3 12B，清楚地体现了斯洛文尼亚语适配带来的益处。GaMS3的表现也超过了规模相近的斯洛文尼亚语专用模型GaMS-9B-Nemotron和Zlatorog。在多语言模型中，表现最佳的是Gemma 3 27B和Bielik-11B。后者主要使用波兰语数据进行训练，因而这一结果令人意外。另一方面，EuroLLM和Apertus本应支持所有欧盟语言，但表现相当糟糕。

表8. 基于各基准测试平均模型排名的斯洛文尼亚语大语言模型评测排行榜。排名越低越好。

模型	平均排名
GaMS-27B-Nemotron	3.05
Gemma 3 27B	3.20
GaMS3-12B	4.25
Bielik-11B	5.00
GaMS-9B-Nemotron	5.20
Zlatorog-12B	5.60
Gemma 3 12B	6.30
Qwen3-30B-A3B	7.30
EuroLLM-22B	7.50
Apertus-8B	7.60

6.2 英语到斯洛文尼亚语翻译

为了评估 GaMS3-12B-Instruct 的英语到斯洛文尼亚语翻译能力，我们使用了自定义基准测试的升级版（Vajda、Vreš 和 Robnik-Šikonja，2025年）。我们通过新增一个数据集和感兴趣的指标对基准测试进行了升级。原始基准测试包含来自 CC-News 和英文维基百科的文章。在此次评估中，我们加入了 GaMS-Nemotron-Chat 数据集中的示例，用于评估 Markdown 格式的保留情况。对于用于聊天数据翻译的模型而言，这类评估必不可少，因为聊天数据采用 Markdown 格式，潜在的格式错误可能会导致对目标模型进行误导性的聊天微调。

与 Vajda、Vreš 和 Robnik-Šikonja（2025年）类似，我们使用简单的 将以下英文文档翻译成斯洛文尼亚语 提示词，将英文文档翻译成斯洛文尼亚语。我们计算了以下指标：用于衡量语义相似度的 COMET、用于检测回复语言错误的语言错误指标，以及用于检测回复过短的截断错误指标。每个数据集的 COMET 分数均使用 CometKiwi 的无参考直接评估（DA）模型（Rei 等，2022年）单独计算。语言检测使用 FastText 库中的预训练语言识别模型（Grave 等，2018年）完成。截断错误通过启发式方法检测。如果译文与原文的字符数之比低于 0.7，则将该示例标记为截断。最后，我们使用 Gemma-3-27b-it 模型（Team，2024 年）测量了存在 Markdown 格式错误的示例比例。该模型接收源文本和译文，并被要求判断两者的格式是否完全匹配（完整提示词见附录 2）。

我们将 GaMS3 与上一代 GaMS 模型、专用 GaMS 翻译器、商用 Gemini-2.5-flash 以及开源 EuroLLM 进行了比较。结果见表 9。尽管参数更少，GaMS3-12B-Instruct 的整体 COMET 性能仍与 GaMS-27B-Instruct 相当。值得一提的是，GaMS-27B-Instruct 在监督微调期间也使用了 RSDO 翻译数据集中的一部分示例进行训练，而 GaMS3 并未使用。GaMS3 的语言错误和格式错误率也显著更低。我们将这一差距归因于

训练数据的局限性。GaMS-27B-Instruct 持续预训练所用的词元数量显著较少（150 亿至 200 亿），对于其规模而言似乎训练不足；而 GaMS3 则受益于使用更多斯洛文尼亚语文本进行持续预训练。此外，GaMS-Nemotron-Chat 并未纳入 GaMS-27B-Instruct 的基础训练，而后者仅使用了少得多的数据（约 2.5 万个示例）进行指令微调，因此出现了更多格式错误。还需要指出的是，GaMS3 使用 GaMS-9B-SFT-DPO-Translator 翻译的维基百科进行预训练，因此在维基百科数据集上的 COMET 分数更高。

与 Gemma 3 相比，GaMS3 在三个数据集上的 COMET 分数都显著更高，这表明斯洛文尼亚语专门化提升了译文的语义相似度。正如预期，GaMS3 的语言错误更少。斯洛文尼亚语专门化的益处也体现在 Markdown 错误的数量上。然而，与两个 Gemma 3 模型相比，GaMS3 的截断错误更多。这表明，未来我们应着重改进长上下文监督微调数据。虽然 GaMS 模型的性能略低于 Gemini，但应注意，Gemini 的规模大得多。此外，Gemini 是领先的翻译商业模型之一，其表现可与 DeepL 等专用系统相媲美。与 EuroLLM 相比，GaMS3 的 COMET 分数更高，错误率也更低，尽管 EuroLLM 使用了大量平行数据进行训练，并且在监督微调阶段也接触了大量翻译示例。

表 9. 英语到斯洛文尼亚语翻译结果。总体 COMET 分数为三个数据集分数的平均值。对于分数越高越好的指标，使用 (↑) 标记；对于分数越低越好的指标，使用 (↓) 标记。每项指标的最高分以粗体显示，第二高的结果带下划线。

	COMET (↑)						
模型	总体	CC-News	Nemotron-Chat	维基百科	语言错误（↓）	截断错误（↓）	Markdown 错误（↓）
Gemini-2.5-flash	**0.717982**	**0.702981**	**0.697498**	**0.753924**	0.35%	0.42%	3.70%
GaMS-9B-SFT-DPO-Translator	0.708042	0.702903	0.679462	0.742583	0.91%	0.28%	18.28%
GaMS-27B-Instruct	0.701284	0.686480	0.680014	0.730733	27.28%	5.36%	62.07%
GaMS3-12B-Instruct	0.700226	0.684848	0.674523	0.741844	1.11%	1.83%	32.61%
GaMS-9B-Instruct	0.693659	0.685006	0.673394	0.723470	13.50%	4.83%	33.15%
EuroLLM-9B-Instruct	0.689321	0.668084	0.670723	0.729227	8.97%	1.89%	35.08%
Gemma-3-27b-it	0.668378	0.643716	0.660861	0.704403	5.93%	**0.0%**	39.17%
Gemma-3-12b-it	0.613083	0.602837	0.619401	0.621018	16.48%	0.19%	57.32%

6.3 Slovene-LLM-Arena

受 LMSYS 等热门聊天机器人竞技场的启发（Chiang 等，2024年；Zheng 等，2024年），我们推出了斯洛文尼亚语大语言模型竞技场，旨在由斯洛文尼亚语用户使用斯洛文尼亚语提示词评估模型。首先，用户需要撰写提示词并获得两个不同模型的回答。模型名称对用户隐藏，模型的回答分别标记为模型 A 和模型 B。用户可以对回答进行排名，也可以继续对话。为回答排名时，用户可以从 4 个选项中选择：模型 A 更好、模型 B 更好、平局，或两个回答都不好。只有在排名完成后，模型名称才会向用户揭晓。排行榜采用源自国际象棋的 ELO 分数 [16] 计算。

我们在竞技场中比较了若干多语言模型和商业模型。表 10 展示了截至2026年1月13日的排行榜快照（排行榜持续更新）。分析结果时，需要注意一些模型的投票数较少，因此 ELO 分数可能会产生误导。不过，将其与胜率结合起来看，仍能提供有用的信息。可以看出，GaMS3 的表现略逊于 GaMS-Nemotron 模型。表现较差的主要原因是，GaMS3 生成的斯洛文尼亚语似乎比另外两个 GaMS 模型生成的内容更

16. https://en.wikipedia.org/wiki/Elo_rating_system

像机器翻译的。这表明我们需要对机器翻译数据集进行筛选和改进，尤其是GaMS-Nemotron-Chat 数据集以及长文本持续预训练阶段使用的数据集。

表 10. 2026年1月13日 Slovene-LLM-Arena 上基于 ELO 分数排名的前 15 个模型排行榜。胜率按获胜次数除以获胜次数与失败次数之和计算（平局不计入）。

Rank	模型	ELO 分数	胜率	总投票数
1	Gemini-2.5-Pro	1100	78.3%	181
2	GaMS-27B-Instruct-Nemotron	1062	66.3%	240
3	Gemma-3-27b-it	1049	66.1%	177
4	GaMS-9B-Instruct-Nemotron	1043	64.2%	129
5	Gemini-2.0-Flash	1036	61.0%	156
6	GPT-4o	1030	59.0%	181
7	GaMS3-12B-Instruct	1025	61.2%	95
8	Kimi-K2-Instruct	1015	58.5%	52
9	Gemini-3-Pro-Preview	1008	83.3%	6
10	DeepSeek-R1	1005	51.2%	166
11	GPT-5	1005	75.0%	5
12	Gemini-1.5-Pro	1004	52.1%	159
13	DeepSeek-R1-0528	1003	51.2%	63
14	GPT-5-Mini	1002	100.0%	1
15	Mistral-Large-3-675B	1000	50.0%	4

与商业模型相比，GaMS3 的表现与 GPT-4o 和 Gemini-2.0-Flash 相当；鉴于这两个模型在规模和资源上的优势，这是一个非常好的结果。所有 GaMS 模型均被 Gemini-2.5-Pro 超越，我们相信，一旦 Gemini-3-Pro 和 GPT-5 获得足够数量的投票，GaMS 模型也会被它们超越。不过，这在意料之中，因为它们是领先模型开发者最新且表现最佳的模型。我们观察到，商业模型在更高难度的任务上，即需要推理的任务上，优于 GaMS3 模型，这表明我们应将这一能力纳入未来的模型中。

与开源模型相比，GaMS3 仅被 Gemma 3 27B（一个非常强大的多语言模型）和 GaMS-Nemotron 模型超越。作为 GaMS3 基础模型的 Gemma 3 12B 未能进入前 15 名，目前仅排第**25**名，ELO 分数为**979**，胜率为 **42.6%**，投票数为 **168**。Zlatorog 是另一个面向斯洛文尼亚语的专门化模型，排名为**第23**，ELO 分数为 **986**，胜率为 **45.2%**，投票数为 **166**。这些结果清楚地表明了我们的斯洛文尼亚语专业化流程所带来的益处。

7. 结论

我们介绍了 GaMS3-12B 的开发过程。这是一款专为斯洛文尼亚语打造的开源大语言模型。我们详细介绍了流程和数据准备，使我们能够以 Gemma 3 12B 模型为基础，对 GaMS3 进行持续预训练和监督微调。我们在三种不同场景下对 GaMS3 模型进行了详细评估：一套斯洛文尼亚语基准测试、Slovenian-LLM-Eval（通用知识）、一项下游生成任务（机器翻译），以及通用聊天性能（斯洛文尼亚语大语言模型竞技场）。

我们方法的主要缺点似乎在于对机器翻译数据的依赖，尤其是在监督微调阶段。因此，该模型的斯洛文尼亚语并不像它本可以达到的那样自然。我们计划在未来通过减少机器翻译数据的数量并提高其质量来解决这一问题。我们的模型有待改进的领域是那些困难任务，无论是在

长上下文和逻辑推理方面。我们计划通过增加长上下文数据的数量，并将推理数据加入我们的监督微调数据集来解决这一问题。此外，尽管 Gemma 3 具备多模态能力，我们仅聚焦于语言模态。我们计划在未来通过与 SVILA（基于 Gemma 3 的斯洛文尼亚语视觉语言模型）模型携手合作来解决这一问题。GaMS3 未来也将接受工具调用方面的训练，从而支持智能体方法。

除了开源模型之外，最重要的贡献是对我们的训练方法和数据准备进行了详细说明。我们证明了，这种方法能够显著提升基础模型在低资源语言中的性能。我们的 GaMS3 模型在所有评测场景中都优于 Gemma 3 12B，并且在斯洛文尼亚语大语言模型竞技场中，其性能可与 GPT-4o 和 Gemini-2.0-Flash 等规模大得多的商业模型相媲美。通过与类似项目开展合作，我们相信，我们的工作能够为其他低资源语言开发类似的开源模型作出重要贡献。

致谢 我们感谢所有参与数据收集和准备工作的人员，正是这些工作使我们能够训练模型。特别感谢 Jaka Čibej 及其团队对基准测试进行人工校正，感谢 Tomaž Savodnik 将维基百科转换为 Markdown 格式并抓取大学知识库，感谢 Miha Malenšek 准备斯洛文尼亚语法律和医疗数据，感谢 Alina Machidon 准备斯洛文尼亚语代码数据集，感谢国家与大学图书馆与我们分享其数字档案，也感谢 Taja Kuzman 和 Nikola Ljubešić 协助进行网络数据整理。

该模型的开发得到了英伟达主权人工智能计划的支持。感谢提供 NVIDIA DGX Cloud Lepton 的使用权限。我们也衷心感谢英伟达一群杰出人士给予的所有支持和帮助：Anna Louise Ollerenshaw、Meriem Bendris、Oleg Sudakov、Benedetta Delfino、Rita Fernandes Neves、Andrea Pilzer、Miguel Martinez、Noel Osagie、Adam Henryk Grzywaczewski 和 Aleks Polak。

部分计 算资源由 SLING 通过项目 S24O01-42 提供 .

参考文献

Apertus 项目、Alejandro Hernández-Cano、Alexander Hägele、Allen Hao Huang、Angelika Romanou、Antoni-Joan Solergibert、Barna Pasztor等。2025年。*Apertus*：让全球语言环境中的大语言模型更加开放、普惠且合规。 https://arxiv.org/abs/2509.14233。

Bañón、Marta、Malina Chichirau、Miquel Esplà-Gomis、Mikel L. Forcada、Aarón Galiano-Jiménez、Cristian García-Romero、Taja Kuzman 等。2023。《斯洛文尼亚语—英语平行语料库 *MaCoCu-sl-en 2.0*。》斯洛文尼亚语语言资源库 CLARIN.SI。http://hdl.handle.net/11356/1813。

Brown、Tom 等。2020年。《语言模型是少样本学习者》。载于《神经信息处理系统进展》33:1877–1901。 https://proceedings.neurips.cc/paper_f iles/paper/2020/f ile/1457c0d6bf cb4967418bf b8ac142f 64a-Paper.pdf。

Chiang、Wei-Lin、Lianmin Zheng、Ying Sheng、Anastasios N. Angelopoulos、Tianle Li、Dacheng Li、Banghua Zhu 等。2024年。《聊天机器人竞技场：一个通过人类偏好评估大语言模型的开放平台》。载于第*41*届国际会议论文集 *f*机器学习会议 *g*htt ps://dl.acm.or g/doi/abs/10.5555/3692070.3692401.

CJVT. 2024年。斯洛文尼亚语大语言模型评测。 https://huggingf ace.co/datasets/cjvt/slovenian-llm-eval。Hugging Face 仓库 .

Comanici, Gheorghe, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein 等。2025年。*Gemini 2.5*：以高级推理、多模态、长上下文和下一代智能体能力突破前沿。 https://arxiv.org/abs/2507.06261。

Cui、Yiming、Ziqing Yang 和 Xin Yao。2023年。面向中文 *LLaMA* 和 *Alpaca* 的高效且有效的文本编码。https: // arxiv.org/abs/2304.08177。

Dubey、Abhimanyu 等。2024年。*Llama 3* 模型家族。https://arxiv.org/abs/2407.21783。

Erjavec, Tomaž. 2023.斯洛文尼亚语混合语料库 *metaFida 1.0*。 斯洛文尼亚语语言资源库 CLARIN.SI。 http://hdl.handle.net/11356/1775。

Gao, Leo, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster等人。2024年。*语言模型评测工具套件*。https://doi.org/10.5281/zenodo.12608602。Gordić, Aleksa。2024年。*Slovenian GPT*。Hugging Face 仓库。https://huggingf ace.co/gordicaleksa/SlovenianGPT。Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, 和 Tomas Mikolov。2018年。为157种语言学习词向量。载于第十一届国际语言资源与评测会议论文集（*LREC 2018*）。https://aclanthology.org/L18-1550/。Guo, Daya, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu 等。2025年。DeepSeek-R1通过强化学习激励大语言模型进行推理。《自然》645（8081）：633–638。http://dx.doi.org/10. 1038/s41586-025-09422-z。Indyk, Piotr, 和 Rajeev Motwani。1998年。近似最近邻：迈向消除维数灾难。载于第三十届 *ACM* 年度计算理论研讨会论文集，604–613。Kosem, Iztok, Jaka Čibej, Kaja Dobrovoljc, Tomaž Erjavec, Nikola Ljubešić, Primož Ponikvar, Mihael Šinkec, 和 Simon Krek。2025年。斯洛文尼亚语监测语料库 *Trendi 2025-07*。斯洛文尼亚语语言资源库 CLARIN.SI。http : //hdl.handle.net/11356/2044。Krek, Simon, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan 等。2021年。*ssj500k 2.3*训练语料库。斯洛文尼亚语语言资源库 CLARIN.SI。http://hdl.handle.net/11356/1434。Kydlíček, Hynek, Guilherme Penedo, 和 Leandro von Werra。2025年。*FinePDFs*。https://huggingf ace.co/datasets/HuggingFace FW/f inepdf s。Hugging Face 仓库。Li, Chong, Shaonan Wang, Jiajun Zhang, 和 Chengqing Zong。2024年。通过跨语言对齐提升多语言生成式语言模型的上下文学习。载于*2024*年北美计算语言学协会年会：人类语言技术论文集（第一卷：长篇论文），8058–8076。https://doi.org/10.18653/v1/2024.naacl-long.445。https://aclanthology.org/2024.naacl-long.445/。Li, Tianle, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, 和 Ion Stoica。2025年。从众包数据到高质量基准测试：arena-hard与benchbuilder流程。载于第四十二届国际机器学习会议。https://openreview.net/f orum?id=Kf Tf 9vFvSn。Ljubešić, Nikola, 和 Taja Kuzman。2024年。CLASSLA-web：经过语言学与体裁标注的南斯拉夫语支语言可比网络语料库。载于*2024*年计算语言学、语言资源与评测联合国际会议论文集（*LREC-COLING 2024*），3271–3282。https://aclanthology.org/2024.lrec-main.291/。Machidon, Alina, 和 Domen Vreš。2025年。斯洛文尼亚代码反馈。https://huggingf ace.co/datasets/cjvt/sl_code_f eedback。Hugging Face 仓库。Mahabadi, Rabeeh Karimi, Sanjeev Satheesh, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, 和 Bryan Catanzaro。2025年。*Nemotron-CC-Math*：*1330*亿词元规模的高质量数学预训练数据集。https://arxiv. org/abs/2508.15096。Martins, Pedro Henrique, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal 等。2025年。EuroLLM：面向欧洲的多语言语言模型。《计算机科学学报》255：53–62。https://www. sciencedirect.com/science/article/pii/S1877050925006210。Meta。2025年。*Llama-4-Maverick-17B-128E-Instruct-FP8*。https://huggingf ace.co/meta-llama/Llama-4-Maverick-17B-128E- Instruct-FP8。Hugging Face 仓库。Mistral-AI, : Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample 等。2025年。*Magistral*。https://arxiv.org/abs/2506.10910。Nanonets。2025年。*Nanonets-OCR-s*。https://huggingf ace.co/nanonets/Nanonets-OCR-s。Hugging Face 仓库。Nathawani, Dhruv, Shuoyang Ding, Vitaly Lavrukhin, Igor Gitman, Somshubra Majumdar, Evelina Bakhturina, Boris Ginsburg, 和 Jane Polak Scowcroft。2025年。*Nemotron-Post-Training-Dataset-v2*。https://huggingf ace.co/datasets/nvidia/Nemotron-Post-Training-Dataset-v2。Hugging Face 仓库。Nathawani, Dhruv, Igor Gitman, Somshubra Majumdar, Evelina Bakhturina, Ameya Sunil Mahabaleshwarkar, Jian Zhang, 和 Jane Polak Scowcroft。2025年。*Nemotron-Post-Training-Dataset-v1*。Hugging Face 仓库。https://huggingf ace.co/datasets/nvidia/Nemotron-Post-Training-Dataset-v1。Nguyen, Thao, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, 和 Xian Li。2025年。网络数据再利用：一种提升语言模型预训练数据质量与数量的方法。载于第二届语言建模会议。https://openreview.net/f orum?id=lkjhBdz3rn。

英伟达。2025年。*NVIDIA Nemotron 3*：高效、开放的智能。https://arxiv.org/abs/2512.20856。英伟达, Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte 等。2025年。*NVIDIA Nemotron Nano 2*：一种准确高效的混合式 *Mamba-Transformer* 推理模型。https : // arxiv.org/abs/2508.14444。Ociepa, Krzysztof, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, 和 Adrian Gwoździej。2025年。*Bielik v3 small*：技术报告。https://arxiv.org/abs/2505.02550。OpenLeecher。2024年。*Lmsys chat 1m clean*。Hugging Face 仓库。https://huggingface.co/datasets/OpenLeecher/lmsys_ chat 1m clean._ _OpenLMLab。2023。*Open-Chinese-LLaMA*。Hugging Face 仓库。https://huggingf ace.co/openlmlab/open-chinese- llama-7b-patch。Penedo, Guilherme, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, 和 Thomas Wolf。2025年。*FineWeb2*：一套流程扩展所有语言——为每种语言调整预训练数据处理。https://arxiv.org/abs/2506.20920。Petrič, Timotej, Domen Vreš, Iztok Lebar Bajec, Tjaša Arčon, Gašper Jelovčan, 和 Marko Robnik-Šikonja。2025a。*GaMS- 27B-Instruct-Nemotron*。https://huggingf ace.co/cjvt/GaMS-27B-Instruct-Nemotron。Hugging Face 仓库。. 2025b。*GaMS-9B-Instruct-Nemotron*。https://huggingf ace.co/cjvt/GaMS-9B-Instruct-Nemotron。Hugging Face 仓库。Petrič, Timotej, Domen Vreš, Iztok Lebar Bajec, 和 Marko Robnik-Šikonja。2025年。*GaMS-Nemotron-Chat* 数据集。Hugging Face 仓库。https://huggingf ace.co/datasets/cjvt/GaMS-Nemotron-Chat。Rei, Ricardo, Marcos Treviso, Nuno M. ~~Guerr~~eiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza 等。2022年。CometKiwi：IST-unbabel 在2022年质量估计共享任务中的参赛成果。载于第七届机器翻译会议论文集（*WMT*），634–645。https://aclanthology.org/2022.wmt-1.60/。

Santos, Rodrigo, João Ricardo Silva, Luís Gomes, João Rodrigues 和 António Branco。2024年。借助 Open Decoder Gervásio PT* 推进葡萄牙语生成式人工智能的发展。载于低资源语言特别兴趣组第三届年度会议论文集（*LREC-COLING 2024*），第 16–26 页。https://aclanthology.org/2024.sigul-1.3。

Savodnik, Tomaž. 2025a. *Wikipedia-Markdown*。 https://huggingf ace.co/datasets/zidsi/wikipedia_ markdown. Hugging Face 仓库。. 2025b. *Zlatorog-12b-instruct-beta*。 Hugging Face 仓库。https://huggingf ace.co/zID4si/Zlatorog-12B-Instruct- Beta. Šorn, Mojca, Ana Cvek, Jure Skubic, Tamara Logar, Sašo Zagoranski, and Alen Bratanović. 2024. 面向大语言模型的斯洛文尼亚语指令数据集 *GaMS-Instruct-DH 1.0*。斯洛文尼亚语语言资源库 CLARIN.SI。http://hdl.handle.net/11356/1975. AI-Sweden. 2024. *GPT-SW3*。 Hugging Face 仓库。https://huggingf ace.co/AI-Sweden-Models/gpt-sw3-40b. Team, Gemma. 2024.*Gemma 2*：以实用规模改进开放语言模型。 https://arxiv.org/abs/2408.00118. . 2025.*Gemma 3* 技术报告。 https://arxiv.org/abs/2503.19786. Thellmann, Klaudia, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, et al. 2024. 面向欧洲语言的多语言大语言模型评测。 https://arxiv.org/abs/2410.08928. Tiedemann, Jörg. 2012. OPUS 中的平行数据、工具与接口。载于第八届国际语言资源与评测会议论文集（*LREC'12*）。~~Ulčar,~~ M., and M. Robnik-Šikonja. 2020. FinEst BERT 和 CroSloEngual BERT：多语言模型中少即是多。载于文本、语音与对话 *TSD 2020*，第 12284 卷。https://doi.org/10.1007/978-3-030-58323-1_11. Ulčar, Matej, and Marko Robnik Šikonja. 2021. SloBERTa：斯洛文尼亚语单语大型预训练掩码语言模型。载于*2021*年信息社会第*24*届国际多学科会议框架下举行的 *SI-KDD* 论文集，17–20. http://library.ijs.si/Stacks/Proceedings/InformationSociety/2021/IS2021_Volume_C.pdf 。Ulčar, Matej, and Marko Robnik-Šikonja. 2023. 面向低资源斯洛文尼亚语的序列到序列预训练。*Frontiers in Artificial Intelligence* 6. https://doi.org/10.3389/f rai.2023.932519. Vajda, Dario, Domen Vreš, and Marko Robnik-Šikonja. 2025. 使用合成偏好数据改进用于机器翻译的大语言模型。载于第二届 *LUHME* 研讨会论文集，67–73。

Vreš, Domen, Iztok Lebar Bajec, Tjaša Arčon, Gašper Jelovčan, and Marko Robnik-Šikonja. 2025. *GaMS-9B-Instruct*。 https://huggingf ace.co/cjvt/GaMS-9B-Instruct. Hugging Face 仓库。Vreš, Domen, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik-Šikonja. 2024. 参数量为 10 亿的低资源语言生成模型。载于语言技术与数字人文会议。 https: //www.sdjt.si/wp/wp-content/uploads/2024/09/JT-DH-2024_Vres_Bozic_ Potocnik_Martincic_Robnik.pdf. Wang, Yizhong, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. 自指令：利用自生成指令对语言模型进行对齐。载于计算语言学协会第*61*届年会论文集（第*1*卷：长篇论文），13484–13508. Wei, Haoran, Yaofeng Sun, and Yukun Li. 2025. *DeepSeek-OCR*：上下文光学压缩。https://arxiv.org/abs/2510. 18234. Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 技术报告，https://arxiv.org/abs/2505.09388. Žagar, Aleš, Matic Kavaš, Marko Robnik-Šikonja, Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Marko Ferme, et al. 2022a. *KAS*语料库摘要 *KAS-Abs 2.0*。 斯洛文尼亚语语言资源库 CLARIN.SI。http://hdl.handle.net/ 11356/1449. . 2022b. 学术斯洛文尼亚语语料库 *KAS 2.0*。 斯洛文尼亚语语言资源库 CLARIN.SI。http://hdl.handle. net/11356/1448. Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, et al. 2024. LMSYS-Chat-1M：大规模真实世界大语言模型对话数据集。载于第十二届国际学习表征会议。 https://openreview.net/f orum?id=BOf DKxf wt0. zID4si. 2024. *TinySLlama-INST*。 Hugging Face 仓库。https://huggingface.co/zID4si/TinySLlama-INST. Zuo, Jingwei, Maksim Velikanov, Ilyas Chahed, Younes Belkada, Dhia Eddine Rhayem, Guillaume Kunsch, Hakim Hacid, et al. 2025.*Falcon-H1*：重新定义效率与性能的混合头语言模型系列。 https://arxiv.org/ abs/2507.22448。

附录 1. 翻译与身份校正提示

本节介绍构建 GaMS-Nemotron-Chat 数据集时使用的提示和模板。

附录 *1.1* 翻译模板

以下提示词模板用于将英语对话翻译为斯洛文尼亚语，同时保留原有的对话结构。

```
### 任务

将下方聊天内容翻译成斯洛文尼亚语。

### 输出

请按以下方式设置回复格式:

<聊天>
<用户>...</用户>
<助手>...</助手>
<用户>...</用户>
...
</聊天>

### 聊天内容

<聊天>
{% for msg in conversation -%}
{% if msg.role   =="用户" -%}
<用户>{{ msg.content }}</用户>
{% elif msg.role   =="assistant" -%}
<助手>{{ msg.content }}</助手>
{% endif -%}
{% endfor -%}
</chat>

### 最后总结

翻译整段对话，并尽可能采用最佳 Markdown 结构。翻译成斯洛文尼亚语。以“<chat>”开头
```

附录 *1.2* 身份纠正提示词

这个提示词用于重新生成模型错误地自我标识（例如称自己为“Qwen”）时的回复，确保模型始终将自己标识为 GaMS。

```
### 指令

你的任务是根据以下上下文回答用户的问题（问题位于 <提示词> 和 </提示词>之间）。请始终依据本




### 你的身份

<身份>
我是 GaMS（斯洛文尼亚语生成模型），一个正在“借助大语言模型进行自然语言适配处理”项目中开发的大









### 问题

<提示>{{ user_


请将回答组织得清晰易读。回答时请结合问题（位于 <poziv> 和 </poziv> 之间）及重要信息，并体
```

附录2. Markdown 格式评估提示

第 6.2 节中用于评估 Markdown 格式的提示词如下。该提示词与 Gemma 3 27B it 搭配使用，以检测给定译文的 Markdown 格式是否与原文完全一致。通过人工检查，我们发现，尽管分类结果并非总是正确，但被判定为不准确的译文总数对我们的研究目的而言足够可靠。

```
你是一名评审，负责检查译文是否保留了原文的格式和风格。


规则
- 仅关注格式和风格，忽略翻译质量和含义。
- 格式和风格包括：标题（使用 # 或下划线）、粗体/斜体/删除线，
行内代码/代码块（围栏数量和语言）、列表（有序/无序）、嵌套以及
列表项数量、块引用、链接（URL 目标必须匹配）、图像（URL 目标必须
匹配）、表格（行数/列数及是否包含表头）、水平分隔线、数学定界符
（$...$、$$...$$）、HTML 标签及其顺序、段落/块边界。
- 段落内的空白差异可以接受。
- **纯文本不是标题。** 标题必须明确以“#”（Markdown）开头，或
使用下划线语法（’---’ 或 ’===’）。
- 结构差异（多出/缺少区块、层级不同、项目数量不同，
网址有变、缺少数学围栏等）都不允许。
- 原文可能是 Markdown 或纯文本。译文必须保持相同的
风格。
- 如果原文仅为纯文本（没有 Markdown、标题、列表、代码围栏，
没有特殊格式），那么 TRANSLATION 也必须是纯文本，才算格式良好。

判定规则
- 只有当所有样式元素和数量都匹配时，才算格式良好。- 否则就是格式不良。


输出（仅限这些标签，不得包含其他内容）<推理>如果有不匹配的元素，请简要总结</reasoning><解释>




原文：<原文开始>




翻译：<翻译文本开始>
```

附录 3. 光学字符识别提示词

本节介绍了使用 Llama-4-Maverick 和 Nanonets 对 PDF 文档逐页进行光学字符识别时所用的提示词。

附录 *3.1 Llama4 OCR* 提示词

我们使用以下斯洛文尼亚语提示词作为 Llama-4-Maverick 的输入。

```
将图像中的文本转换为 Markdown 文本格式。如果图像不包含文本，则返回空文档。转换时请严格遵循以
```

提示词的英语翻译：

```
将图像中的文本转换为 Markdown 文本格式。如果图像中没有文本，则返回空文档。转换时，请严格遵守

- 注意正确书写数学方程、常数和变量。符号应采用 LaTeX 格式；同时，注意方程是位于文本行中还是单

- 仅返回图像中的文本，不要返回任何其他内容。- 注意图像中文字的写法。转录内容应与图像完全一致，不

- 注意不要包含页眉和页脚。有些页面会在页眉中写有书籍或杂志的标题，并在页脚中标注页码。请省略这

- 图像中的文本可能是旧报纸的一部分。请注意文本的结构。
```

附录*3.2 Nanonets OCR* 提示词

我们使用 将以下提示词作为输入提供给 Nanonet；它是模型模型卡的一部分。 s:

```
像自然阅读一样，从上述文档中提取文本。请以 HTML 格式返回表格。请以 LaTeX 表示方程。如果文档中
```