跨语言的 Text2Cypher：评估基础模型超越英语的能力

Makbule Gulcin Ozsoy
Neo4j
London, UK
makbule.ozsoy@neo4j.com

William Tai
Neo4j
London, UK
william.tai@neo4j.com

摘要

大型语言模型的最新进展使自然语言接口能够将用户的问题翻译成数据库查询，例如 Text2SQL、Text2SPARQL 和 Text2Cypher。虽然这些接口增强了数据库的可访问性，但目前大多数研究仅限于英语，并且在其他语言上的评估有限。本文探讨了基础 LLM 在多种语言下的 Text2Cypher 任务表现。我们创建并发布了一个多语言测试集，通过将英文问题翻译成西班牙语和土耳其语，同时保留原始的 Cypher 查询，从而实现公平的语言间比较。我们使用标准化提示和指标对多个基础模型进行了评估。我们的结果显示了一致的表现模式：英语最好，其次是西班牙语，最差的是土耳其语。我们将此归因于训练数据可用性的差异以及语言特征的不同。此外，我们还探索了将任务提示翻译成西班牙语和土耳其语的影响。结果表明，评价指标几乎没有变化，这表明提示的翻译影响较小。我们的研究强调了在多语言查询生成中需要更加包容的评估和发展。未来的工作包括模式本地化以及跨多种语言进行微调。

PVLDB 参考格式：
Makbule Gulcin Ozsoy and William Tai. 跨语言的 Text2Cypher：评估基础模型超越英语的能力。PVLDB, 14 (1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

终端用户授权书，卷。14 ，编号。1 ISSN 2150-8097。
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The translated dataset is available at HuggingFace: https://huggingface.co/datasets/mgoNeo4j/translated_text2cypher24_testset

1 介绍

数据库提供了高效的数据存储、组织和检索机制。查询语言如 SQL（用于关系型数据库）、SPARQL（用于 RDF 图）或 Cypher（用于图数据库）使用户能够与这些系统进行交互 [7]。大型语言模型 (LLMs) 的进步使得可以将自然语言问题翻译成数据库查询（Text2SQL，Text2SPARQL，Text2Cypher），从而使数据库对非专家更易于访问。然而，该领域的大多数研究都集中在英语上，而很少关注其他语言 [6, 8]。

这项工作专注于 Text2Cypher，将自然语言问题转换为可执行的 Cypher 查询（见图 1）。例如，用户可能希望为问题*汤姆·汉克斯的电影有哪些？*编写一个 Cypher 查询，该问题可以用不同的语言表达，如英语（EN）、西班牙语（ES）或土耳其语（TR）。无论使用哪种语言，Text2Cypher 模型都应生成相同的查询，如*"MATCH (演员:人 {姓名: "Tom Hanks"})-[:出演]->(电影:电影) RETURN 电影.标题"*。此外，我们旨在评估基础 LLM 在这一任务上的多语言性能。为此，我们构建了一个测试集，其中包含英语、西班牙语和土耳其语的语义对齐的问题，并且所有问题均与同

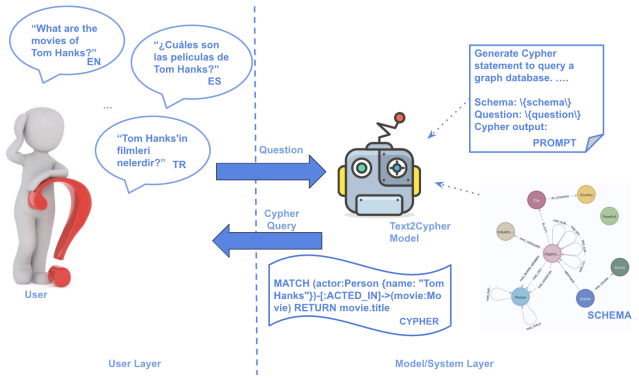


图 1: 用户希望为问题“汤姆·汉克斯的电影有哪些”生成一个 Cypher 查询。然而，他们可能会用不同的语言表达同样的问题，例如英语（EN）、西班牙语（ES）或土耳其语（TR）。

一个 Cypher 查询配对。这种设置能够进行跨语言基准测试，从而将输入语言的影响与模型性能隔离。我们的主要贡献是：

- 我们发布了 Text2Cypher 测试集的多语言版本 [15]，将其翻译成西班牙语和土耳其语，以实现跨语言评估 [1]。这些语言因其语言多样性和资源差异而被选中。
- 我们使用标准化的提示和指标在此数据集上评估了多个基础模型，并通过语义等价的问题比较了英语、西班牙语和土耳其语之间的性能。
- 我们分析了语言对 Text2Cypher 性能的影响。我们的结果显示，模型在英语上的表现最好，其次是西班牙语，在土耳其语上表现最差。我们将土耳其语较低的表现归因于其粘着结构和与印欧语系的语言距离。

2 相关工作

大型语言模型（LLMs）已成为广泛自然语言处理任务的核心。虽然许多早期的 LLMs 主要在英语数据上进行训练，但最近通过开发多语言模型 [13] 扩展了它们的能力。研究人员也开始探索这些模型的内部机制 [12, 17–19] 并评估它们在各种应用领域中的跨语言性能 [8, 14]。多语言模型是在包含多种语言的数据集上进行训练的，使它们能够学习跨语言表示。这使得它们可以利用从高资源语言（如英语）中学到的词关联和语法规则，并将其应用于有较少可用训练数据的语言。然而，由于英语仍然主导着大多数训练集，这些模型也可能将特定于英语的语法模式和假设转移到其他语言环境中 [13, 18]。因此，在高资源语言及其相似语言中，LLMs 的表现显著优于低资源语言 [13]。

在将自然语言翻译成数据库查询语言的背景下，许多多语言工作集中在 Text2SQL 任务上 [4, 9]。最近的研究考虑了诸如葡萄牙语 [8, 16]、阿拉伯语 [1]、俄语 [2] 和土耳其语 [11] 等语言。在这项工作中，我们研究了 Text2Cypher 任务，该任务将自然语言问题翻译成 Cypher 查询。

3 TEXT2CYPHER 超越英语

3.1 数据集准备

我们使用了公开可用的 Text2Cypher 数据集 [15]，它包括英语问题、数据库模式、真实 Cypher 查询和元数据。为了评估跨语言的基础模型，我们使用 LLM 将测试集的问题字段翻译成西班牙语和土耳其语，并比较所有三种语言的结果。英语是一种资源极其丰富的印欧语系语言，西班牙语是一种资源丰富的印欧语系语言，而土耳其语是一种资源中等的阿尔泰语系语言，具有独特的语言特征 [10, 13]。基于此，我们预计西班牙语的性能会略有下降，而土耳其语的性能会大幅下降，这是由于语言差异造成的。我们使用了以下翻译步骤：

- **屏蔽命名实体和引用**：为了确保命名实体和引文在翻译过程中保持不变，我们首先将它们进行了遮罩。命名实体被替换为指示其类型和索引的占位符（例如，LOCATION_0），而引文字符串则被替换成 QUOTE_<index>标记。例如，句子 “Hello, I work at ’Neo4j’ in London” 被遮罩为 “Hello, I work at QUOTE_0 in LOCATION_0”。

[1] 数据集:https://huggingface.co/datasets/mgoNeo4j/translated_text2cypher24_testset

表 1: {语言名称}的翻译提示

翻译指令提示
您是一位专业的翻译人员。您的任务是将给定的文本从英语翻译成 {language_name}。遵循以下指南: 1. 保持原文的意义和语气 2. 保留任何占位符格式 <TYPE_NUMBER>（例如，<PERSON_0>, <QUOTE_1>） 3. 使翻译在 {language_name} 中自然流畅 4. 保持正确的格式和标点符号 5. 不翻译或修改任何占位符 6. 不翻译引号内的任何文本 7. 不翻译命名实体（专有名词、名称、地点、组织） 8. 不翻译数字、日期或度量单位 请仅提供翻译后的文本，不要有任何解释或额外的上下文。

- **使用大语言模型的翻译** ：我们使用 GPT-4o-mini 模型将带掩码的问题进行了翻译，参照了表 1 中所示的提示。
- **恢复掩码**：翻译后，我们将占位符替换为其原始命名实体和引述表达式以重构最终问题。

请注意，真实的 Cypher 查询语句没有被翻译。这些查询包含（i）特定于 Cypher 的语法（例如，匹配，其中），（ii）模式中的术语（例如，人物，出演于，电影），以及（iii）用户提供的字面量（例如，汤姆 · 汉克斯），所有这些都是为了保持其原始形式。

3.2 实验设置和评估指标

我们在 Text2Cypher 数据集的测试部分进行实验 [15]。为了支持多语言评估，我们使用一个大语言模型将原始英文问题自动翻译成西班牙语和土耳其语。在翻译过程中，有 50 个样本存在屏蔽相关的问题并被排除在外。因此，每种语言都在剩下的 4,783 个样本上进行了评估。我们使用的基础模型如下：(i)**Gemma-2-9b-it:** 发布于 2024 年 6 月。主要支持英语。(ii) **元 llama-3.1-8b-instruct:** 发布于 2024 年 7 月。主要支持英语和几种欧洲语言。(iii)**Qwen2.5-7B-Instruct:** 发布于 2024 年 9 月。主要支持中文和英文，并且保持对超过 29 种语言的多语言支持。我们使用它们的 Unsloth 版本模型，这些模型经过 Instruct 微调并量化为 4 位精度（例如 unsloth/gemma-2-9b-it-bnb-4bit），以提高效率和易用性。对于 Text2Cypher 任务，我们使用与先前工作相同的提示 [15]，如表 2 所示。生成后，会有一个额外的后处理步骤用于删除不需要的文本，例如 'cypher:' 后缀。我们使用 HuggingFace Evaluate 库 [5] 来计算评估指标。我们采用两种评估程序：(i) **基于翻译的：** 根据文本内容将生成的 Cypher 查询与参考查询进行比较。我们报告此评估的 ROUGE-L 分数。(ii) **基于执行的：** 在目标数据库上执行生成和参考查询，并比较它们的输出（按字典顺序排序）。我们报告此评估的 Exact-Match 分数。

4 实验结果

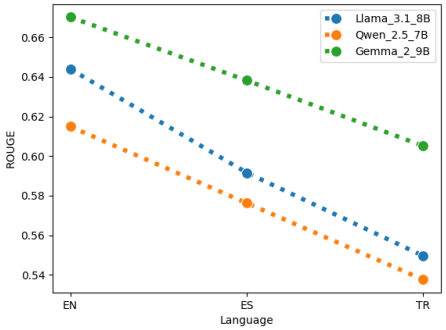
我们在英文、西班牙文和土耳其文中使用 Text2Cypher 对基础 LLM 进行评估，重点在于问题和提示语言。

4.1 问题语言的影响

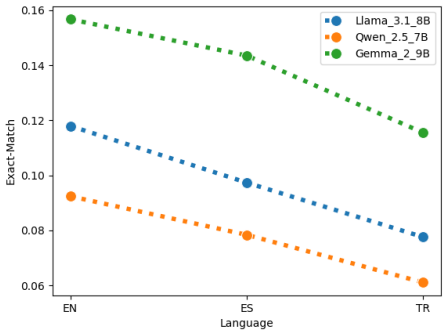
我们首先使用翻译后的数据集来考察问题语言对基础 LLM 的影响。在这个阶段，我们不对 Text2Cypher 任务的任务提示进行修改（如表 2 所示），该提示保持为英语。图 2 显示了基础模型在英语、西班牙语和土耳其语问题上的表现。结果显示所有模型在英语问题上表现最佳，其次是西班牙语，在土耳其语问题上的表现最低。这种模式与语言资源丰富度的差异相吻合：英语是一种极高的资源语言，西班牙语是高资源语言，而土耳其语则是中等资源语言 [10, 13]。LLM 通常在训练数据更丰富和更多样的语言上表现更好。此外，先前的研究 [3] 显示模型跨语言相似性更强的语言时泛化效果更好。由于西班牙语和英语都属于印欧语系，而土耳其语属于阿尔泰语系，结构上的差异可能进一步导致了性能的差异。在这些实验中，提示保

表 2: 用于 Text2Cypher 任务的指令

类型	指令提示
System Instruct.	Task: Generate Cypher statement to query a graph database. Instructions: Use only the provided relationship types and properties in the schema. Do not use any other relationship types or properties that are not provided in the schema. Do not include any explanations or apologies in your responses. Do not respond to any questions that might ask anything else than for you to construct a Cypher statement. Do not include any text except the generated Cypher statement.
User Instruct.	Generate Cypher statement to query a graph database. Use only the provided relationship types and properties in the schema. Schema: {schema} \n Question: {question} \n Cypher output:



(a) 基于翻译的评估：ROUGE-L 分数



(b) 基于执行的评估：完全匹配分数

图 2: 模型在 Text2Cypher 上的表现当输入问题是英语（EN）、西班牙语（ES）或土耳其语（TR）。

持为英语，我们认为这可能会引入当处理非英语问题时对 LLM 造成的困惑。在下一节中，我们通过展示一个基础模型使用翻译后的问题和翻译后的提示的结果来探索这一点。

4.2 提示语言的影响

我们还将提示翻译成西班牙语和土耳其语，以评估提示语言对模型性能的影响。这些提示使用 GPT-4o-mini 模型进行了翻译，并展示在表 3 中。带有和不带翻译提示的西班牙语和土耳其语的性能对比结果呈现在图 3 中。结果显示，当提示被翻译时，用于基于翻译评估真实文本与生成文本差异的 ROUGE-L 分数略有提高（约 1 – 1.5%）。然而，对于基于执行评估的 Exact-Match 分数则基本保持不变。这些结果表明，在我们的设置中，提示翻译对整体结果的影响较小。值得注意的是，在我们的数据集中，数据库模式，包括节点、关系和属性，仍然使用英语。未来的研究可以探索将模式项本地化以匹配输入和提示语言的影响，这可能揭示语言与性能之间的进一步交互。

5 结论

本工作展示了基础 LLM 在 Text2Cypher 任务上，英语、西班牙语和土耳其语之间的性能比较。我们创建并发布了一个多语言测试集，通过将英语问题翻译成西班牙语和土耳其语，同时保持相同的基准查询，实现了公平的跨语言模型性能对比。我们的结果显示，基础模型在英语问题上的表现最佳，其次是西班牙语，而在土耳其语上表现最差。这种模式可能是由于训练

表 3: 西班牙语（ES）和土耳其语（TR）用于 Text2Cypher 任务的指令

类型	指令提示（ES）	指令提示（TR）
System Instruct.	Tarea: Generar una sentencia Cypher para consultar una base de datos de grafos. Instrucciones: Usa únicamente los tipos de relaciones y propiedades proporcionados en el esquema. No utilices ningún otro tipo de relación ni propiedades que no estén incluidas en el esquema. No incluyas explicaciones ni disculpas en tus respuestas. No respondas a ninguna pregunta que no sea una solicitud para construir una sentencia Cypher. No incluyas ningún texto excepto la sentencia Cypher generada.	Görev: Bir çizge veritabanını sorgulamak için bir Cypher ifadesi oluştur. Talimatlar: Yalnızca şemada verilen ilişki türlerini ve özellikleri kullan. Şemada verilmeyen herhangi bir ilişki türünü veya özelliği kullanma. Yanıtlarında hiçbir açıklama veya özür ifadesine yer verme. Cypher ifadesi oluşturmak dışında başka bir şey isteyen sorulara yanıt verme. Oluşturulan Cypher ifadesi dışında hiçbir metin ekleme.
User Instruct.	Genera una sentencia Cypher para consultar una base de datos de grafos. Usa únicamente los tipos de relaciones y propiedades proporcionados en el esquema. Esquema: {schema} \n Pregunta: {question} \n Salida Cypher:	Bir çizge veritabanını sorgulamak için Cypher ifadesi oluştur. Şemada verilen ilişki türleri ve özellikler dışında hiçbir şeyi kullanma. Şema: {schema} \n Soru: {question} \n Cypher çıktısı:

数据可用性的差异以及语言特征的不同所驱动，这些因素为模型带来了额外的挑战。我们还探讨了任务指令提示与输入问题一起翻译的影响。研究发现表明，提示翻译稍微改善了基于翻译的评估结果，但对查询执行结果几乎没有影响，这表明在我们的设置中，提示语的语言对整体性能影响较小。未来的工作包括模式本地化以提高多语言性能、针对个别或组合语言进行微调以及扩展到更多语言的评估。

REFERENCES

[1] Saleh Almohaimeed, Saad Almohaimeed, Mansour Al Ghanim, and Liqiang Wang. 2024. Ar-Spider: Text-to-SQL in Arabic. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. 1024–1030.

[2] Daria Bakshandaeva, Oleg Somov, Ekaterina Dmitrieva, Vera Davydova, and Elena Tutubalina. 2022. PAUQ: Text-to-SQL in Russian. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2355–2376.

[3] Tejas Indulal Dhamecha, Rudra Murthy V, Samarth Bharadwaj, Karthik Sankaranarayanan, and Pushpak Bhattacharyya. 2021. Role of language relatedness in multilingual fine-tuning of language models: A case study in indo-aryan languages. *arXiv preprint arXiv:2109.10534* (2021).

[4] Longxu Dou, Yan Gao, Mingyang Pan, Dingzirui Wang, Wanxiang Che, Dechen Zhan, and Jian-Guang Lou. 2023. MultiSpider: towards benchmarking multilingual text-to-SQL semantic parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 12745–12753.

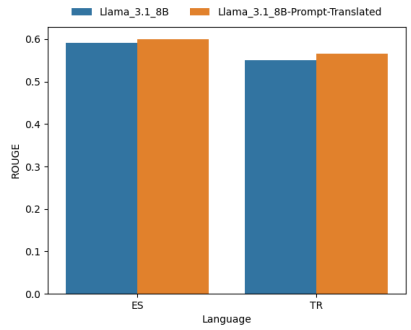
[5] HuggingFace Evaluate. 2024. https://huggingface.co/evaluate-metric.

[6] Xiang Geng, Ming Zhu, Jiahuan Li, Zhejian Lai, Wei Zou, Shuaijie She, Jiaxin Guo, Xiaofeng Zhao, Yinglu Li, Yuang Li, et al. 2024. Why Not Transform Chat Large Language Models to Non-English? *arXiv preprint arXiv:2405.13923* (2024).

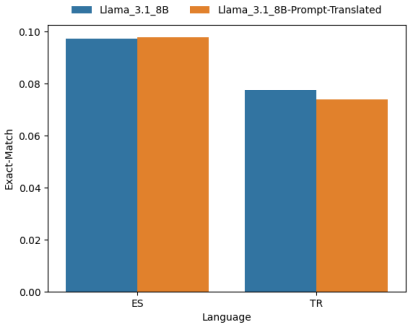
[7] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d' Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)* 54, 4 (2021), 1–37.

[8] Marcelo Jannuzzi, Yuriy Perezhohin, Fernando Peres, Mauro Castelli, and Aleš Popovič. 2024. Zero-Shot Prompting Strategies for Table Question Answering with a Low-Resource Language. *Emerging Science Journal* 8, 5 (2024), 2003–2022.

[9] Marcelo Archanjo José and Fabio Gagliardi Cozman. 2021. mRAT-SQL+ GAP: a Portuguese text-to-SQL transformer. In *Intelligent*



(a) 基于翻译的评估：ROUGE-L 分数



(b) 基于执行的评估：精确匹配分数

图 3: 模型性能（有/无）提示翻译。

Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10. Springer, 511–525.

[10] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095* (2020).

[11] Ali Bugra Kanburoglu and Faik Boray Tek. 2024. TURSpider: A Turkish Text-to-SQL Dataset and LLM-Based Study. *IEEE Access* (2024).

[12] Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2024. MEXA: Multilingual Evaluation of English-Centric LLMs via Cross-Lingual Alignment. *arXiv preprint arXiv:2410.05873* (2024).

[13] Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: large language models in non-English content analysis. *arXiv preprint arXiv:2306.07377* (2023).

[14] Makbule Gulcin Ozsoy. 2024. Multilingual Prompts in LLM-Based Recommenders: Performance Across Languages. *arXiv preprint arXiv:2409.07604* (2024).

[15] Makbule Gulcin Ozsoy, Leila Messallem, Jon Besga, and Gianandrea Minneci. 2025. Text2Cypher: Bridging Natural Language and Graph Databases. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*. 100–108.

[16] Breno Carvalho Pedroso, Marluce Rodrigues Pereira, and Denilson Alves Pereira. 2025. Performance evaluation of LLMs in the Text-to-SQL task in Portuguese. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*. SBC, 260–269.

[17] Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do Multilingual LLMs Think In English? *arXiv preprint arXiv:2502.15603* (2025).

[18] Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055* (2024).

[19] Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in? *arXiv preprint arXiv:2408.10811* (2024).