A Comparative Analysis of Large Language Models for Bilingual Term Extraction in Spanish-Arabic Interpreting and Translation

Mahmoud Gaber
mahmoudgaber@uma.es
IUITLM, University of Malaga
29/10/2025, TRADITUR 2025

Table of contents

01

Introduction

A Brief Overview of the Study

Introduction

Large Language Models (LLMs) are transforming Natural Language Processing (NLP), with growing interest in their role in terminology extraction. As a key aspect of translation and interpreting, terminology extraction ensures semantic precision in specialised communication. Automatic Terminology Extraction (ATE) seeks to ease manual term management by generating ranked term lists from domain corpora. However, empirical research on LLMs for ATE—especially for linguistically distant pairs like Spanish-Arabic—remains limited. This study compares four AI tools (ChatGPT, DeepSeek, Gemini, and Manus) for bilingual term extraction in ophthalmology and tourism. Using Precision, Recall, F-score, and Accuracy, it offers insights into AI's effectiveness in specialised, cross-linguistic contexts.

02

Rationale

Rationale and Motivation

Justification and Motivation

Why Focus on Underrepresented Languages?

- *Dominant Focus: AI-based terminology extraction research is predominantly centred on major European languages.*
- *Critical Gap: This leaves a significant blind spot regarding performance for underrepresented language pairs.*

The Urgency: LLMs in the Wild

- *Large Language Models (LLMs) are being rapidly integrated into training and professional translation settings.*
- *We must proactively assess their reliability across diverse linguistic contexts before their use becomes entrenched.*

The Stakes: Why Terminology Matters

- *Accurate bilingual terminology is the cornerstone of quality in specialised translation and interpreting (Corpas Pastor, 2018; Sales, 2024).*

Justification and Motivation

The Shortcomings of Traditional Methods

Current corpus-based approaches face two major issues:

- *The Data Problem: Scarcity: Extreme difficulty in obtaining suitable parallel corpora (Daille 2012; Delpech et al. 2012).*

- *The Technical Problem: Alignment: Persistent challenges in achieving accurate sentence and term alignment (Castillo Rodríguez 2011; Gaber 2025).*

03

Objectives

Research objectives

Research Objectives

- **Benchmark** the performance of ChatGPT-4o, DeepSeek, Gemini, and Manus.
- **Select** the most suitable tool for Spanish-Arabic term extraction.
- **Outline** pathways for improving AI-assisted terminology management.

04

RQs

Research Questions

Research Questions

- **Effectiveness**: How reliable are LLMs for Spanish-Arabic bilingual term extraction?
- **Performance**: Which AI tool scores highest on Precision, Recall, F-score, and Accuracy?
- **Improvement**: What are the main limitations and key areas for enhancement?

05

Methodology

Methodological Approach

Methodology

- **Dataset** and Corpus Selection:

Domain : Medical (Ophthalmology), Cultural Tourism
Corpus **type** : Comparable Corpus
Languages : (Spanish-Arabic)

Corpus name	Size
Tour_Cor_AR	2.021
Tour_Cor_ES	2.105
Ophtal_Cor_AR	8.023
Ophtal_Cor_ES	1.812

Methodology

Gold set (human-curated Gold Set of 75 terms) in Spanish

- **AI tools** Selected: **ChatGPT**-4o**, DeepSeek**-R1**, Copilot, Gemini and Manus**
- **Prompting** Techniques**:** mixed approach (CREATE and CO-STAR)
  - **Terminology extraction**
  - **Equivalents extraction and suggestion**

<Prompt for terminology extraction (AR) Tourism>

Role:
You are a seasoned terminologist with 20+ years of experience in **tourism** terminology, specialising in **cultural tourism**.

Context:
You are tasked with extracting **Arabic** terminology from a provided text for translation and interpreting purposes. The text is related to **cultural tourism**.

Instructions:

1. **Scope:**
    - Extract **single-word and multi-word terms** (including abbreviations/acronyms) directly related to **cultural tourism** from the attached file.
    - **Only include terms explicitly present in the text** (do not infer or invent terms).
    - Prioritize terms with high relevance to cultural tourism.
2. **Output Requirements:**
    - Format: **Excel-compatible table**.
    - Columns:
        - **Term (Arabic) | Notes (context from text)**
    - Extract **as many terms as possible** without sacrificing accuracy.
3. **Quality Checks:**
    - Avoid duplicates.
    - Verify that terms are cultural tourism-specific.

Methodology

- **Evaluation Metrics**

Compare **4 term lists** (extracted by 4 different LLMs) against a **human-curated Gold Set** and compute:

Precision = [Correctly Extracted Terms] / [Total Terms Extracted by LLM]
Recall = [Correctly Extracted Terms] / [Total Terms in Gold Set]
F-score = 2 * (Precision * Recall) / (Precision + Recall)
Accuracy = [Correct Terms] / [Total Terms in Gold Set + Incorrect LLM Terms]
Jaccard Index = [Gold Set ∩ LLM Terms] / [Gold Set ∪ LLM Terms]

06

Rs & D

Results and Discussion

Performance Analysis Results I Tour_COR_ES

Model Name	Precision (%)	Recall (%)	F-score (%)	Accuracy (%)	Jaccard Index (%)
ChatGPT-4o	84.91	46.15	59.70	38.30	38.30
DeepSeek-R1	88.14	64.10	74.26	54.90	54.90
Gemini	86.27	58.97	70.14	49.32	49.32
Manus	82.61	51.28	63.16	42.11	42.11

Analysis-I

DeepSeek performs best across all metrics with the highest Precision (88.14%), Recall (64.10%), F-score (74.26%), Accuracy (54.90%), and Jaccard Index (54.90%)

Precision is generally high across all models (82-88%), indicating they're good at avoiding incorrect terms

Recall shows more variation, with DeepSeek capturing the most relevant terms from the Gold Set

ChatGPT has good precision but lower recall, suggesting it's conservative in term extraction

Gemini shows balanced performance, second-best overall

Manus has the lowest precision and moderate recall

Performance Analysis Results II Ophtal_COR_ES

Model Name	Precision (%)	Recall (%)	F-score (%)	Accuracy (%)	Jaccard Index (%)
ChatGPT-4o	84.44	50.67	63.33	46.34	46.34
DeepSeek-R1	87.50	46.67	60.87	43.75	43.75
Gemini	45.26	57.33	50.59	33.86	33.86
Manus	54.88	60.00	57.32	40.18	40.18

Analysis-II

A clear **trade-off between precision and recall** emerged across models. Models with **higher precision** (DeepSeek, ChatGPT) tended to have **lower recall**, while models with higher recall (Manus, Gemini) demonstrated lower precision.

Conservative extractors: DeepSeek and ChatGPT appear to be more selective, prioritizing correctness over comprehensiveness.
Liberal extractors: Manus and Gemini seem to cast a wider net, capturing more gold standard terms but including more incorrect terms.

Analysis -II

This evaluation reveals significant variations in how different LLMs approach terminology extraction in specialized domains. While ***DeepSeek and ChatGPT demonstrated superior precision, and Manus showed stronger recall****,* ***no model achieved excellent performance across all metrics****. The* ***highest F-score (ChatGPT's 63.33%) still falls below*** *what would be considered strong performance in most applications*

Conclusion

These findings suggest that **current LLM-based terminology extraction systems have substantial room for improvement**, particularly in achieving better balance between precision and recall.

Common Issues in Terminology Extraction Using Large Language Models (LLMs)

- Limited identification of **multi-word terms**.
- Difficulty recognising compound terms whose components do not appear together.
- Inconsistent recognition and extraction of **acronyms**.
- Redundancy and **repetition** of extracted terms.
- Inclusion of items that do **not** qualify as **valid terms**.

Thanks!

🙋‍♂️ Q&A + Open Discussion

mahmoudgaber@uma.es

References

Castillo Rodríguez, Cristina. 2011. “La Alineación de un Corpus Paralelo Multilingüe: Propuesta de Fases para la Didáctica de Traducción Especializada Inversa.” Cadernos de Tradução 27 (1): 117–142. https://doi.org/10.5007/2175-7968.2011v1n27p117

Corpas Pastor, Gloria, and Lily May Fern. 2016. "A Survey of Interpreters' Needs and Practices Related to Language Technology." Technical Paper FFI2012-38881-MINECO/TI-DT-2016-1.

Jaccard, Paul (1901). "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". *Bulletin de la Société vaudoise des sciences naturelles (in French). 37(142): 547–579.*

Sales, Dora. 2024. “Professional Translators' and Interpreters' Views on Information Competence: An Exploratory Qualitative Study from the Spanish Context.” Journal of Librarianship and Information Science 56 (3): 743–59. https://doi.org/10.1177/09610006231164l

Veisbergs, Andrejs. 2006. “Dictionaries and Interpreters." EURALEX Proceedings. 1219–1224. https://www.euralex.org/elx_proceedings/Euralex2006/146_2006_V2_Andrejs%20VEISBERGS_Dictionaries%20and%20Interpreters.pdf

Acknowledgment:

This work was carried out in the framework of the following research projects: Postdoctoral research contract (PPIT-UMA), PIE22-135 (2022/23-2023/24), VIP II (PID2020-112818GB-I00/AEI/10.13039/501100011033), RECOVER (ProyExcel_00540), DIFARMA (HUM106-G-FEDER), and DÍGAME (JA.A1.3-06).