Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 79-94 recibido 04-12-2025 revisado 19-01-2026 aceptado 12-02-2026

Generative query parsing and multilingual semantic content retrieval in scientific domain

Segmentación de consultas multilingües y recuperación semántica de contenidos para el ámbito científico

Nicolau Duran-Silva[1,2], **Pablo Accuosto**[1], **Horacio Saggion**[2]
[1]SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain
[2]LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain
{nicolau.duransilva,pablo.accuosto}@sirisacademic.com, horacio.saggion@upf.edu

Abstract: Recent multilingual large language models enable richer access to scientific information, but their use in low- and mid-resource settings remains underexplored. We evaluate whether domain-adapted generative and embedding models can improve multilingual scientific information retrieval across Catalan, Spanish, and English. Two tasks are addressed: multilingual query parsing and cross-lingual semantic search through adapted sentence embeddings. Our results show that compact multilingual models, when tuned with domain-specific research data, provide accurate and language-agnostic access to open research information.
Keywords: Multilingual semantic retrieval, query segmentation, scientific information access, domain adaptation.

Resumen: Los grandes modelos del lenguaje multilingües permiten un acceso más completo a la información científica, pero su uso en entornos con recursos disponibles limitados permanece poco explorado. Evaluamos si los modelos generativos y de incrustación adaptados al dominio pueden mejorar la recuperación de información científica multilingüe en catalán, español e inglés. Abordamos los problemas de segmentación de consultas multilingües y la búsqueda semántica croslingue mediante incrustaciones de frases adaptadas. Nuestros resultados muestran que los modelos multilingües compactos, cuando se ajustan con datos de investigación multilingue proporcionan un acceso, preciso e independiente de idioma, a la información de investigación abierta.
Palabras clave: Recuperación semántica multilingüe, segmentación de consultas, acceso a información científica, adaptación a dominios especificos.

1 Introduction

Scientific and technical knowledge is increasingly available through open databases of research projects, publications, and patents (Lin et al., 2023). Local and international funding agencies—such as the Generalitat de Catalunya (Fuster et al., 2023)[1], AEI[2] CDTI[3], and the European Commission's CORDIS platform[4]—publish large amounts of structured and textual data describing funded projects, their participants, and outcomes. This expanding landscape of *open research information* is inherently multilingual (Céspedes et al., 2025), and in our target ecosystem, most records are written in Catalan, Spanish, or English. Records vary widely in terminology, metadata quality, and the level of descriptive detail in their textual content (e.g., some include full abstracts while others contain only titles). These characteristics make information access challenging for researchers and policymakers who need to explore, compare, and interpret research and innovation (R&I) activities across languages and institutions. While traditional keyword-based information retrieval (IR) systems can handle metadata fields, they fail to capture the semantic relationships and contextual meaning that characterize modern scientific language—for example, they cannot recognize that 'oncology' and 'cancer research' refer to related concepts, or that a query in Spanish should match an English abstract on the same topic.

[1]https://ris3mcat.gencat.cat/
[2]https://www.aei.gob.es/ayudas-concedidas
[3]https://www.cdti.es/
[4]https://cordis.europa.eu/projects

ISSN 1135-5948 DOI 10.26342/2026-76-6

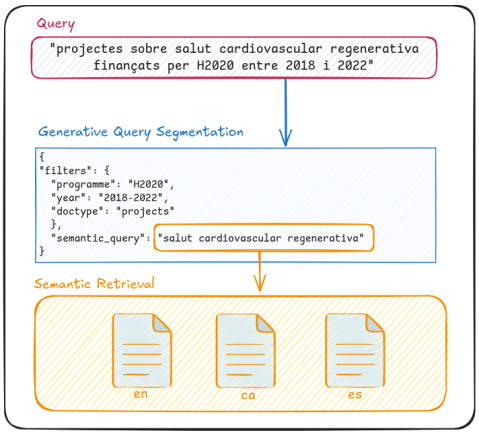

Figure 1: Overview of the pipeline and subtasks.

Recent advances in multilingual large language models (LLMs) have enabled more natural and concept-oriented interaction with textual databases through semantic search (Biswal et al., 2024). Embedding-based retrieval represent search queries and document in shared semantic space, supporting retrieval beyond exact word matches. Nevertheless, R&I information access often requires a hybrid approach: users must search both textual content (e.g., abstracts) and structured metadata (e.g., programmes, years, organizations), combining semantic understanding with precise filtering. Pure embedding-based document retrieval does not address this compositional nature of user intent. Generative models (Tola, 2024) offer a promising solution by converting natural-language queries into structured forms aligned with database schemas while mitigating hallucinations through retrieval grounding.

In this context, we investigate how domain-adapted multilingual language models can improve access to scientific and innovation data, with a particular focus on the Catalan–Spanish–English setting, which characterizes most R&I records in our target ecosystem. Our goal is to translate user queries into structured filters consistent with R&I metadata schemas (Mosca, Roda, and Rull, 2018), which models entities such as projects, organizations, funding programmes, and temporal attributes. At the same time, we explore techniques for injecting scientific knowledge and enhancing multilingual alignment in embedding models to support cross-lingual and concept-driven semantic search. An overview of the problem we address is presented in Figure 1. This study aims to contribute to the evaluation of the *AINA* project.[5] While AINA models (Gonzalez-Agirre, 2025) have shown competitive performance on general NLP benchmarks, their suitability for real-world scientific IR tasks remains underexplored. To address these gaps, our work investigates two complementary directions:

1. **Use and fine-tuning of multilingual generative models for query parsing:** converting natural language queries into structured JSON filters aligned with the UNICS R&I schema (Mosca, Roda, and Rull, 2018), with a focus on open-source models.
2. **Scientific knowledge injection and multilingual alignment in sentence-embedding models:** improving semantic retrieval across Catalan, Spanish, and English, particularly for cross-lingual and concept-centric search scenarios.

Together, these experiments assess whether small multilingual generative and embedding-based models can enable accurate, language-agnostic access to structured scientific information. We release models, datasets, and code to support further research in multilingual scientific information access.[6]

2 *Background*

Accessing scientific information in multilingual R&I ecosystems requires two complementary capabilities: (a) interpreting multilingual natural queries, (b) retrieving related scientific and technical document across languages. These correspond to two well-established NLP tasks, semantic parsing of queries and dense document retrieval, both of which have evolved substantially with the advent of large language models.

2.1 Query Understanding and Structured Parsing

Translating natural language queries into structured database operations has evolved from early rule-based systems (Green Jr et al., 1961; Woods, 1972; Warren and Pereira,

[5] An effort led by the Barcelona Supercomputing Center to develop open multilingual natural language processing (NLP) resources and foundational models for Catalan and other co-official languages in Spain.

[6] `https://github.com/sirisacademic/aina-impulse.git`

1982) through statistical and neural approaches (Zelle and Mooney, 1996; Zettlemoyer and Collins, 2007; Dong and Lapata, 2016) to modern LLM-based methods, extensively reviewed in recent works (Liu and Xu, 2025; Hong et al., 2025). The emergence of LLMs has fundamentally transformed text-to-SQL parsing, enabling strong zero-shot and few-shot performance (Hong et al., 2025; Mohammadjafari, Maida, and Gottumukkala, 2024). Nonetheless, generating syntactically valid structured outputs remains challenging, motivating schema-aware prompting, retrieval-augmented in-context learning, and parameter-efficient fine-tuning (Dettmers et al., 2023)). Recent work on structured generation, including StructuredRAG (Shorten et al., 2024), shows that complex JSON prediction requires explicit schema conditioning. Beyond relational settings, Text-to-NoSQL has recently gained attention with benchmarks such as TEND (Lu et al., 2025) and MultiTEND (Qin et al., 2025), enabling text-to-query mapping for document-oriented databases. Prior efforts include LLaMA-based fine-tuning for NoSQL (Tola, 2024) and schema discovery from natural language (Mior, 2024).

Query parsing techniques have evolved from unified mention extraction and linking approaches (Ma et al., 2020) to joint models for slot filling and intent detection using capsule networks (Zhang et al., 2019) that capture hierarchical relationships between query components. Semantic decomposition methods have proven particularly effective, with (Eyal et al., 2023) introducing Query Plan Language for systematically breaking complex queries into simple subqueries, and SUBS (Yang, Zhang, and Yang, 2022) achieving compositional generalization through subtree substitution. Recent modular frameworks (Wang et al., 2022) combine Named Entity Recognition, Neural Entity Linking, and Neural Semantic Parsing, establishing architectural patterns where specialized components handle different aspects of query understanding, directly relevant to filter extraction in systems like ours. These approaches provide the methodological foundation for our structured JSON query parsing.

Cross-lingual semantic parsing for multilingual support employs sophisticated strategies beyond simple translation, with the Rex framework (Shi et al., 2022a), which interpolates representations from source and translated utterances (representation mixup) to leverage English training data for non-English queries, achieving state-of-the-art on Chinese and Vietnamese benchmarks. Zero-shot approaches (Sherborne and Lapata, 2022) demonstrate successful transfer to multiple languages using only English-logical form pairs and target language corpora, while XRICL (Shi et al., 2022b) explores retrieval-augmented in-context learning for cross-lingual scenarios. These methods are particularly relevant for systems like ours requiring Catalan, Spanish, and English support, suggesting that maintaining semantic consistency across languages requires architectural adaptations beyond simple translation. Addressing data scarcity in specialized domains, recent work explores template-based synthetic generation with (Tran and Tan, 2020) extracting templates using BART for hierarchical semantic parsing, while (Wang et al., 2021) combine non-neural PCFG for program composition with neural translation for natural language generation. Domain adaptation techniques complement synthetic approaches, with (Li et al., 2020) proposing two-stage coarse-to-fine frameworks using domain discrimination and relevance attention. These methods prove essential for specialized applications like our target database, where domain-specific terminology and query patterns differ significantly from general benchmarks, enabling effective model specialization with limited target domain data.

2.2 Semantic Similarity Search and Dense Retrieval

Semantic similarity search is dominated by dense retrieval methods, which encode queries and documents into a shared embedding space and rank candidates by vector similarity (Karpukhin et al., 2020; Izacard et al., 2021). These approaches enable concept-level search beyond lexical overlap, but also face known limitations (Weller et al., 2025) in capturing rare terminology and fine-grained distinctions, issues especially relevant for scientific and multilingual contexts. The sentence-transformers framework (Reimers and Gurevych, 2019) provides a widely adopted pipeline for training dense retrievers, typically using Multiple Negatives Ranking Loss (MNRL) (Hadsell,

Chopra, and LeCun, 2006), where in-batch examples act as implicit negatives to efficiently learn discriminative representations. Recent advances refine contrastive objectives through better negative sampling, hard negative mining (Xiong et al., 2020), crosslingual pairs (Feng et al., 2022), and improved optimization strategies (Singh et al., 2023; Jørgensen and Breitung, 2025), all contributing to more robust vectorial representations.

In the scientific domain, several model families are particularly relevant. The multilingual E5 models (Wang et al., 2024) show strong cross-lingual transfer from large-scale retrieval corpora, while SPECTER (Cohan et al., 2020) leverages citation networks to specialize embeddings for scientific papers (though predominantly in English). Multilingual RoBERTa-based encoders trained in trilingual query relevance dataset (on 65k CA-ES-EN query-passage pairs) demonstrates effectivenes when trained with domain-appropriate data achieves (Rodriguez-Penagos et al., 2021). These models benefit substantially from domain-specific contrastive fine-tuning, which improves discrimination between closely related scientific concepts. Hybrid retrieval architectures partially address the weaknesses of dense-only methods in handling exact matches and rare entities (e.g., uncommon organization names, specialized technical terms, or newly coined concepts that appear infrequently in training corpora). Dense retrieval also plays a central role in retrieval-augmented generation (RAG) frameworks, improving factual accuracy for LLMs (Lewis et al., 2020; Borgeaud et al., 2022). However, adapting dense retrievers to specialized multilingual scientific domains remains challenging due to domain-specific terminology, code-switching, and limited non-English training data (Zhang et al., 2021; Litschko, Vulić, and Glavaš, 2022). Our approach follows the contrastive MNRL paradigm while introducing domain-specific multilingual pairs to strengthen semantic alignment across Catalan, Spanish, and English research texts.

3 Materials and Methods

Our approach consists of two main tasks: query parsing and document retrieval. In this section we describe the datasets, and models developed for each task.

3.1 Query Parsing

The goal of the query parsing task is to automatically transform natural-language queries in Catalan, Spanish, and English into a structured JSON representation aligned with the UNICS schema (Mosca, Roda, and Rull, 2018) for R&I information. Each parsed query contains: (i) a `semantic_query` capturing the main research topic, (ii) a set of `filters` (programme, year, organization, etc.), and (iii) `metadata` such as language or normalization notes.

3.1.1 Datasets

We built a multilingual training set consisting of 682 synthetic queries programmatically generated by combining controlled vocabularies (funding programmes, organization names, locations) with thematic keywords extracted from project titles and abstracts, following domain-specific templates that mirror realistic user queries. The evaluation set contains 100 real queries manually annotated with their corresponding gold JSON structures. Table 1 summarises both datasets. All training samples follow a ChatML-style format[7] with explicit system–user–assistant turns, introduced with instructions and some examples. Examples of annotated queries and their gold segmentation output are provided in Appendix A.

Dataset	Size	Lang.	Description
Train	682	ca/es/en	Template-generated queries; 88% *Discover*, 12% *Quantify*.
Test	100	ca/es/en	Expert-written queries with manually curated gold JSON annotations.

Table 1: Datasets for the query parsing task.

Training Data

The synthetic data were generated from controlled vocabularies and project metadata from RIS3CAT and related R&I sources (Fuster et al., 2023). Queries were programmatically composed using domain-specific templates covering different user intents, resolvability and query types, and component distribution. Detailed information about data generation, query intents, schema resolvability, component distributions, and illustrative examples are provided in Appendix A.

[7]Documentation available at `https://huggingface.co/docs/transformers/v5.0.0rc0/chat_templating`

Evaluation Data
The evaluation set consists of 100 queries elicited from researchers and R&I policy analysts through a structured questionnaire asking them to formulate typical information needs. Each query was manually annotated with a gold-standard JSON structure containing the correct filters, semantic query, and metadata. This dataset is used to assess both component-level accuracy and overall structured-output correctness. However, the dataset is not balanced across languages. English, Catalan, and Spanish queries were independently received from users, resulting in different distributions across languages. In particular, Spanish queries represent a 21% and exhibit higher structural complexity on average, with more filter components and a higher proportion of partially resolvable intents.

3.1.2 Models & Trainings

We fine-tuned six multilingual instruction-tuned LLMs spanning both Catalan-focused models from the AINA project (Salamandra-2B and Salamandra-7B (Gonzalez-Agirre, 2025)) and strong general-purpose multilingual baselines (Mistral-7B (Jiang et al., 2023), Qwen2.5-3B and Qwen2.5-7B (Yang et al., 2025)). This selection enables comparing domain-adapted Catalan models with widely used cross-lingual architectures under the same task. All models share a transformer architecture and were fine-tuned using the same Low-Rank Adaptation (LoRA) setup (Hu et al., 2022) for parameter-efficient fine-tuning. LoRA adapters train only 1% of model parameters while keeping base weights frozen, enabling consistent fine-tuning across models of different sizes. Training was carried out for three epochs on the synthetic multilingual dataset with deterministic low-temperature decoding (0.1) to encourage stable JSON generation (full configuration in Appendix C).

3.2 Embedding Retrieval

The second task focuses on multilingual semantic retrieval, aiming to improve cross-language access to research project descriptions and scientific abstracts. While the query parsing module produces structured filters, this component addresses the complementary problem of representing scientific texts through multilingual sentence embeddings. Our objective is twofold: (i) to improve multilingual alignment, so that a query in one language retrieves relevant documents regardless of their language, and (ii) to inject scientific-domain knowledge and terminology into embedding models. Together, these enable robust semantic search across Catalan, Spanish, and English.

Split	# Pairs	%
Train	61,083	80.0
Dev	7,665	10.0
Test	7,618	10.0
Language Distribution		
Catalan (ca)		33.4
Spanish (es)	-	33.3
English (en)	-	33.3

Table 2: Multilingual query–passage retrieval dataset.

3.2.1 Datasets

We use two datasets for training and evaluating multilingual scientific retrieval models: (a) an automatically constructed multilingual pair dataset for contrastive learning, and (b) a text classification dataset to evaluate class separation.

A. Multilingual Query–Passage Pair Dataset. This is the main dataset used to train multilingual retrieval models, constructed to support monolingual and crosslingual scientific retrieval across Catalan, Spanish and English [8]. Each instance pairs a short query (keyword or title) with a semantically or lexically related scientific text (title, abstract or title+abstract). The dataset contains 76k pairs across train/dev/test splits (80/10/10), summarized in Table 2. More detail about dataset construction and examples of query–text pairs are provided in Appendix B.

B. Document Classification Dataset. To complement the contrastive retrieval dataset with additional scientific-domain knowledge , we also use for evaluation a multilingual text classification dataset derived from `scidocs-mag` (Cohan et al., 2020). The original corpus contains ˜19k scientific publications annotated with 19 disciplinary categories (e.g., *computer vision*, *biology*, *materials science*). Each instance consists of an abstract and its corresponding research field label. To obtain a trilingual version, we translated one third of the abstracts into Catalan and one third into Spanish using Google

[8] Available at `https://huggingface.co/datasets/nicolauduran45/multilingual_research_pairs`

Translate, keeping the remaining third in the original English. The result[9] is a balanced CA–ES–EN classification dataset that supports experiments on domain adaptation and multilingual alignment for scientific content.

3.2.2 Models & Trainings

We fine-tuned multilingual embedding models using a contrastive learning objective to align semantically related texts across languages. The training followed the *Multiple Negatives Ranking Loss (MNRL)* (Hadsell, Chopra, and LeCun, 2006) paradigm implemented in the `sentence-transformers` framework (Reimers and Gurevych, 2019), optimizing for cosine similarity between positive pairs.

Base Models. The following multilingual and domain-adapted base encoders were evaluated:

- `mRoBERTA_retrieval`[10]: trilingual RoBERTa model pre-trained on CA, ES, and EN data.
- `Multilingual E5`[11] (Wang et al., 2024): multilingual text embedding encoder trainer from MS-MARCO, a large-scale passage retrieval dataset derived from Bing search queries. (Bajaj et al., 2016)
- `distilRoBERTa`[12] (Reimers and Gurevych, 2019): lightweight English baseline.
- `SPECTER`[13] (Cohan et al., 2020): scientific-domain encoder trained on English documents for citation similarity, providing strong baseline for semantic paper retrieval.

Training Setup. Fine-tuning was performed on our multingual query-passage pair dataset using `SentenceTransformerTrainer` with the configuration described in Appendix D

4 Results

In this section we report the results obtained in our experiments and description of evaluation metrics.

4.1 Evaluation Metrics

4.1.1 Query Parsing

Evaluation focuses on two aspects: the syntactic correctness of the generated structured output and the semantic accuracy of the extracted components. We focus *verifier-based* and *structured-output* dimensions, with the following metrics:

- **Strict Accuracy**: proportion of queries with fully correct predicted representation, requiring all filters, component fields, the semantic query to match the gold annotation, and the output language to match the input language
- **Relaxed Accuracy**: proportion of queries for which the semantic content is correctly captured, requiring all filters, organisation names, and the semantic query to be correct, while allowing organisation type mismatches and language differences.
- **Language Match**: proportion of queries for which the generated output is in the same language as the input query, determined using a lightweight rule-based language detector applied to the `semantic_query` and `query_rewrite` fields.
- **JSON Validity (%)**: proportion of outputs that are valid JSON and correctly match the required structure (i.e., no parsing errors, all expected keys present).
- **Component Accuracy**: proportion of cases in which individual fields extracted by the model match the gold annotations. We report accuracy for *programme* (e.g., `H2020`, `FEDER`); *time range* (e.g., "`>=2020`", "`2015–2018`"), *location*, *organisation*, *semantic query* (core topic or reformulated of the natural language search intent).

4.1.2 Content Retrieval

The resulting models were evaluated on the held-out multilingual test split using:

- **Cosine Similarity**: average cosine similarity between positive query–passage pairs, measuring embedding alignment quality.
- **Mean Reciprocal Rank (MRR)**: evaluates the ranking quality by measur-

[9] Our dataset is available at `https://huggingface.co/datasets/nicolauduran45/multilingual-research-classification`

[10] `https://huggingface.co/langtech-innovation/mRoBERTA_retrieval`

[11] `https://huggingface.co/intfloat/multilingual-e5-base`

[12] `https://huggingface.co/sentence-transformers/all-distilroberta-v1`

[13] `https://huggingface.co/sentence-transformers/allenai-specter`

ing the inverse rank of the first correct passage within the top-10 retrieved candidates.

- **Top-k Recall** ($k \in \{1, 5, 10\}$): proportion of queries for which the paired passage appears among the top-k retrieved results.
- **Neighbourhood Polarity**: following (Jørgensen and Breitung, 2025) formulation, we compute the proportion of the top-k nearest neighbours in the embedding space that share the same class label (discipline) as the target document. Higher polarity indicates more coherent semantic neighbourhoods and stronger clustering of scientific topics.

4.2 Generative Query Parsing

We evaluate how multilingual LLMs convert natural-language queries into structured JSON aligned with the UNICS schema. Table 3 reports overall parsing quality (strict/relaxed accuracy, JSON validity, language match) comparing base and finetuned models. Fine-tuning achieves consistent improvements across all metrics, with strict and relaxed accuracy increasing by a factor of 2–3× relative to base models. The best-performing configuration (Salamandra-7B, fine-tuned) obtains 51% strict accuracy and 65% relaxed accuracy. Language match between input and output also improves in most configurations. We report accuracy per schema component (programme, year, location, organisations, semantic query) in Table 4, and present, in Table 5, results by query language (EN, CA, ES) to assess multilingual robustness. Together, these results characterise model performance under multilingual and schema-constrained query parsing.

4.3 Embedding Retrieval

Table 6 presents the overall retrieval performance of the base and fine-tuned embedding models. We report R@1, R@5, R@10, and MRR, computed on our multilingual query–passage dataset. To ensure robust evaluation given the semi-supervised construction of the dataset, retrieval is performed over batches of 64 candidate documents, which reduces false positives arising from semantically related but non-paired samples. Within each batch, we also treat as valid positives any passages associated

Model	SA	RA	VJ	LM
Base Models				
salamandra-7b	.15	.29	1	.53
qwen2.5-7b	.19	.44	1	.48
mistral-7b	.25	.29	.93	.74
qwen2.5-3b	.11	.30	.99	.47
salamandra-2b	.00	.00	.15	.03
Fine-tuned Models				
salamandra-7b	.51	.65	1	.87
qwen2.5-7b	.47	.65	1	.85
mistral-7b	.24	.55	1	.59
qwen2.5-3b	.39	.59	1	.80
salamandra-2b	.00	.01	.05	.05

Table 3: Comparative evaluation of fine-tuned models on the multilingual query parsing. SA=Strict Accuracy; RA=Relaxed Accuracy; VJ=Valid JSON; LM=Lang Match.

Model	Prog.	Year	Loc.	Org.	Sem.Q
Base Models					
salamandra-7b	.96	.99	.81	.68	.44
qwen2.5-7b	.96	.99	.93	.75	.67
mistral-7b	.80	.83	.78	.59	.59
qwen2.5-3b	.93	.96	.90	.70	.45
salamandra-2b	.10	.05	.11	.05	.04
Fine-tuned Models					
salamandra-7b	.96	.98	.91	.77	.86
qwen2.5-7b	.97	.99	.92	.82	.81
mistral-7b	.97	.99	.94	.81	.66
qwen2.5-3b	.95	.98	.93	.72	.86
salamandra-2b	.05	.04	.03	.02	.04

Table 4: Component-level accuracy of fine-tuned models on the multilingual query parsing task. Prog=Programme; Org.=Organisation; Sem.Q=Semantic Query.

with additional author keywords from the same source document, reflecting the many-to-many nature of scientific keywords, where multiple terms can describe the same underlying concept and a single document may be relevant to multiple query terms. We complement these metrics with a polarity score derived from the classification dataset, measuring whether the top-k neighbours (here, $k = 16$) share the same scientific field. This provides an external estimate of whether fine-tuned models preserve meaningful semantic structure across scientific domains.

4.3.1 Performance by Query Type

We further evaluate retrieval behaviour by distinguishing whether the query term appears explicitly in the target passage. We automatically label a pair as *lexical* when the query string appears verbatim in the passage (e.g., query "cancer" and passage containing "breast cancer"); and as *semantic* otherwise, when retrieval success requires conceptual inference or paraphrasing (e.g., query "cancer" and passage mentioning "basal cell carcinoma"). This classification is language-

Model	EN	CA	ES
Base Models			
salamandra-7b	.30	.27	.29
qwen2.5-7b	.30	.58	.52
mistral-7b	.35	.21	.29
Fine-tuned Models			
salamandra-7b	.72	.64	.52
qwen2.5-7b	.70	.58	.68
mistral-7b	.59	.45	.62

Table 5: Relaxed parsing accuracy by query language (EN, CA, ES) for the best performing models. With number of samples EN (46), CA (33), and ES (21).

Model	R@1	R@5	R@10	MRR	Polarity
Base Models					
E5	.47	.70	.80	.58	.61
mRoBERTA	.34	.59	.71	.46	.66
DistilRoBERTa	.39	.58	.68	.49	.59
Specter	.27	.47	.59	.37	.57
Fine-tuned					
E5	.74	.91	.95	.81	.71
mRoBERTA	.65	.86	.91	.74	.69
DistilRoBERTa	.62	.81	.88	.71	.65
Specter	.61	.83	.89	.70	.65

Table 6: Multilingual semantic retrieval performance across base and fine-tuned embedding models. Metrics: Recall@k (R@1/5/10), MRR, and neighbourhood polarity at k=16.

agnostic: a pair can be lexical or semantic regardless of whether the query and passage share the same language. For instance, a cross-lingual pair may still be lexical if the translated forms match directly. Table 7 reports Recall@1 and Recall@10 across lexical and semantic matches. We observe a gap between lexical and semantic retrieval, with semantic queries being more challenging across all models. Fine-tuning substantially improves performance for both match types, but the gap remains, indicating that capturing deeper conceptual similarity is still harder than exploiting surface lexical overlap with embedding representations.

4.3.2 Retrieval Results by Language Configuration

To analyse the impact of multilingual setting on retrieval quality, we evaluate the models separately on *monolingual* and *cross-lingual* pairs. The monolingual scenario corresponds to queries and passages written in the same language, while the cross-lingual scenario contains pairs where the query and the target text are in different languages (Catalan, English, Spanish). This distinction allows us to measure both in-language semantic retrieval and the ability of models to align concepts across languages. Table 8 reports Recall@1 and Recall@10 under both

Model	R@1		R@10	
Base Models	**Lex.**	**Sem.**	**Lex.**	**Sem.**
E5	.42	.29	.75	.63
mRoBERTA	.23	.21	.59	.56
DistilRoBERTa	.33	.24	.60	.52
Specter	.17	.16	.44	.43
Fine-tuned Models				
E5	.73	.57	.94	.88
mRoBERTA	.62	.46	.90	.82
DistilRoBERTa	.60	.43	.86	.76
Specter	.56	.43	.86	.78

Table 7: Lexical vs. semantic Recall@1 and Recall@10 for base and fine-tuned models.

Model	R@1		R@10	
Base Models	**Mono.**	**Cross.**	**Mono.**	**Cross.**
E5	.46	.23	.82	.55
mRoBERTA	.25	.19	.60	.55
DistilRoBERTa	.35	.21	.64	.48
Specter	.20	.12	.49	.38
Fine-tuned Models				
E5	.69	.60	.92	.89
mRoBERTA	.58	.49	.87	.84
DistilRoBERTa	.58	.44	.84	.78
Specter	.55	.44	.84	.79

Table 8: Recall@1 and Recall@10 segmented by pair type (monolingual vs. cross-lingual), comparing base and fine-tuned models. *M = monolingual* | *C = cross-lingual*.

conditions for all base and fine-tuned models. Results indicate that monolingual retrieval consistently outperforms cross-lingual retrieval for all models, reflecting a tendency of multilingual encoders to prioritise documents written in the same language as the query. Fine-tuning reduces this gap, which is crucial for our setting, although a monolingual bias remains.

4.3.3 Monolingual Retrieval Performance by Language

We further analyse monolingual retrieval performance by grouping test-set pairs according to the language of the target passage. This allows us to examine how models handle scientific text in each language independently, isolating retrieval accuracy from cross-lingual alignment effects. Table 9 reports Recall@1 and Recall@10 for all models across the three languages. Before adaptation, performance is uneven across languages, with English yielding higher recall than Catalan and Spanish. After fine-tuning, retrieval quality can be more balanced across the three languages, although English has advantage, reflecting its stronger representation in pretraining corpora and predominance in science.

Model	R@1			R@10		
Base Models	**CA**	**EN**	**ES**	**CA**	**EN**	**ES**
E5	.37	.55	.45	.77	.86	.83
mRoBERTA	.19	.30	.26	.55	.65	.59
DistilRoBERTa	.25	.53	.26	.57	.78	.56
Specter	.15	.28	.17	.40	.61	.44
Fine-tuned Models						
E5	.63	.74	.69	.91	.94	.91
mRoBERTA	.53	.60	.61	.85	.88	.88
DistilRoBERTa	.50	.67	.56	.80	.90	.81
Specter	.48	.64	.52	.81	.90	.82

Table 9: Monolingual Recall@1 and Recall@10 across CA/EN/ES pairs.

5 Discussion

The results highlight complementary strengths and limitations of the generative and embedding-search approaches explored in this work. While the two tasks are different both reveal the impact of multilingual domain adaptation and the particular challenges posed by the Catalan–Spanish–English setting.

5.1 Query Parsing

Fine-tuning yields 2–3× improvements in strict accuracy compared to base models, for most of the models. The best-performing model (Salamandra-7B) achieves 51% strict accuracy and 65% relaxed accuracy, demonstrating that compact multilingual generative models can reliably map natural-language queries into structured JSON. The three model families exhibit distinct behaviours: Salamandra shows balanced performance across all three languages, Qwen performs better on English and Spanish than Catalan, while Mistral underperforms considerably in Catalan and Spanish. At the 2-3B scale, only Qwen remains competitive with its 7B counterpart (39% vs. 47% strict). Language consistency—whether the model responds in the same language as the input—is robust for English (95–98%) and Catalan (91–97%) but weaker for Spanish (28–52%), despite balanced training data (33% per language). This pattern persists in base models, suggesting an inherent difficulty in language identification rather than a fine-tuning artefact. Recent research indicates that LLMs process tasks through English-dominant internal representations before rendering outputs in target languages, a process that may fail during structured generation where English syntax dominates training corpora (Artemova and others, 2025; Wendler and others, 2025). Spanish, as a high-resource language, may be more susceptible to this drift due to the "curse of multilinguality"—competing directly with English for model capacity—whereas lower-resource languages like Catalan may occupy more distinctive regions in the embedding space (Conneau et al., 2020). The IberBench evaluation confirms that language identification remains challenging for LLMs across Iberian languages (González et al., 2025). Language match is analysed as a secondary, usability-oriented metric because generated representations are consumed by downstream user-facing components (e.g., faceted searches or query rewrites), even when underlaying retrieval operated cross-lingually. The language-specific results, however, should be interpreted with caution. As described in Section 3.1.1, the evaluation set is not balanced across languages and exhibits differences in query complexity, particularly in Spanish. These confounds make it difficult to isolate language-specific effects from query difficulty. A controlled evaluation using parallel queries would be necessary to draw definitive conclusions about cross-lingual performance. Since multilingual robustness was not the primary focus of this work, we leave such systematic comparison for future research. Although not designed as a controlled ablation, the comparison between Catalan-centric models (Salamandra) and general-purpose multilingual models (Qwen, Mistral) provides insight into the role of language-specific adaptation. While base Catalan models place behind larger general-purpose models, fine-tuning largely closes this gap and can cause more balanced performance across languages.

5.2 Embedding retrieval

Across all experiments, fine-tuning consistently improves retrieval quality for every model and evaluation setting. The gains are especially pronounced in cross-lingual retrieval, where base encoders struggle to align Catalan, Spanish, and English scientific content. Models such as E5 and mRoBERTA show 20–30 point improvements in Recall@1 after contrastive fine-tuning, confirming that domain-specific multilingual alignment is essential for scientific search. These improvements extend across match types: while lexical queries are naturally easier, fine-tuning

boosts semantic retrieval capacity almost as much, demonstrating that the models learn to generalise beyond surface forms. And that models can also improve their sparse retrieval capabilities. Fine-tuning transforms even weaker baselines (Distil-RoBERTa, Specter) into competitive multilingual retrievers. When examining monolingual performance by language, we observe clear asymmetries that reflect underlying resource availability. English remains the easiest setting, with the highest scores even before adaptation. Catalan and Spanish lag behind in base models, particularly Catalan, which suffers from limited representation in pretraining corpora. After fine-tuning, however, these gaps narrow substantially: Catalan gains the largest relative improvements, and Spanish reaches parity with English in some models. Taken together, these results show that the combination of multilingual contrastive learning and modest domain-specific supervision yields robust multilingual and crosslingual semantic search capabilities—crucial for accessing R&I information in ecosystems where English, Spanish, and Catalan coexist.

5.3 Error analysis

For query parsing, we report examples of good, partial and wrong parsing in Appendix E. Exploring several error examples, we observe the most frequent issue is language drift, where the structured content is correct but the generated output is in a different language than the input. This behaviour is especially common for short or keyword-based or short queries. A second error category concerns under-extraction of explicit filters, particularly programmes, which are sometimes treated as semantic-query terms rather than structured constraints. We also observe semantic over-generalisation, where qualifiers are omitted, and organisation-related errors (missing types or incomplete normalisation), which are more frequent in smaller models. Overall, most errors derive from schema grounding, multilingual generation effects, or underspecified input rather than a failure to capture user intent.

For retrieval, a dominant error type is the lexical–semantic gap, performance degrades when retrieval requires conceptual inference rather than exact term overlap, as shown in Table 7. Errors are more frequent in cross-lingual settings (Table 8), reflecting imperfect alignment of multilingual scientific terminology. Additionally, short queries (1–2 tokens) provides substantially lower recall than longer ones (>4 tokens), due to limited contextual information, which could probably be improved with context expansion.

These errors suggest clear paths for improvement. Both tasks would benefit from a deeper investigation into the limitations of synthetic and semi-supervised training data, including potential biases and coverage gaps, as well as from controlled evaluations using fully parallel multilingual benchmarks to better isolate cross-lingual effects.

6 Conclusions

In this work, we investigated how multilingual language models can improve access to scientific and innovation data in a trilingual setting characteristic of many R&I information systems. We introduced two complementary tasks: multilingual query parsing, where generative models translate natural-language queries into structured filters aligned with the UNICS schema, and multilingual semantic retrieval, where embedding models are adapted to support accurate monolingual and cross-lingual search. Our results show that lightweight, domain-adapted models, including Catalan-centric variants, can achieve competitive performance when fine-tuned on R&I data. We observe that fine-tuning substantially improves JSON-structured query interpretation and delivers large gains in cross-lingual and semantic retrieval quality for under-resourced languages such as Catalan. In particular, our evaluation of Catalan-centric models shows that, while they are less competitive than general-purpose multilingual models in their base configuration, domain-specific fine-tuning enables them to reach comparable performance while maintaining more balanced behaviour across Catalan, Spanish, and English. Beyond findings, we contribute new multilingual datasets, model checkpoints, and evaluation resources designed to support future research on scientific information access. Overall, our study highlights the value of domain-specific adaptation and multilingual alignment for enabling more effective access to open research information.

Acknowledgments

This work was supported by the Industrial Doctorates Plan of the Departament de Recerca i Universitats de la Generalitat de Catalunya (grant reference 2022/DI /00017). This work was also supported by the Torres y Quevedo Programme (PTQ) funded by the Agencia Estatal de Investigación (AEI), Government of Spain (grant reference PTQ2022-012416). Additional co-funding was provided by the AINA Challenge, funded by the Barcelona Supercomputing Center and the Generalitat de Catalunya. The views and opinions expressed are those of the authors and do not necessarily reflect those of the funding bodies.

References

Artemova, E. et al. 2025. The translation barrier hypothesis: Multilingual generation with large language models suffers from implicit translation failure. *arXiv preprint arXiv:2506.22724*.

Bajaj, P., D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, et al. 2016. Ms-marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Biswal, A., L. Patel, S. Jha, A. Kamsetty, S. Liu, J. E. Gonzalez, C. Guestrin, and M. Zaharia. 2024. Text2sql is not enough: Unifying ai and databases with tag. *arXiv preprint arXiv:2408.14717*.

Borgeaud, S., A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Céspedes, L., D. Kozlowski, C. Pradier, M. H. Sainte-Marie, N. S. Shokida, P. Benz, C. Poitras, A. B. Ninkov, S. Ebrahimy, P. Ayeni, et al. 2025. Evaluating the linguistic coverage of openalex: An assessment of metadata accuracy and completeness. *Journal of the Association for Information Science and Technology*, 76(6):884–895.

Cohan, A., S. Feldman, I. Beltagy, D. Downey, and D. S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Dong, L. and M. Lapata. 2016. Language to logical form with neural attention. *arXiv preprint arXiv:1601.01280*.

Eyal, B., M. Mahabi, O. Haroche, A. Bachar, and M. Elhadad. 2023. Semantic decomposition of question and sql for text-to-sql parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13629–13645.

Feng, F., Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Fuster, E., T. Fernández, H. Carretero, N. Duran-Silva, R. Guixé, J. Pujol, B. Rondelli, G. Rull, M. Cortijo, and M. Romagosa. 2023. Towards building a monitoring platform for a challenge-oriented smart specialisation with ris3-mcat. *arXiv preprint arXiv:2401.10900*.

González, J. Á., I. Borrego-Obrador, Á. Romo Herrero, A. M. Sarvazyan, M. Chinea-Rios, A. Basile, and M. Franco-Salvador. 2025. Iberbench: Llm evaluation on iberian languages. *arXiv preprint arXiv:2504.16921*.

Gonzalez-Agirre, A. 2025. Marc pamies, joan llop, irene baucells, severino da dalt, daniel tamayo, josé javier saiz, ferran espuna, jaume prats, javier aula-blasco, et al. *Salamandra technical report*.

Green Jr, B. F., A. K. Wolf, C. Chomsky, and K. Laughery. 1961. Baseball: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pages 219–224.

Hadsell, R., S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Hong, Z., Z. Yuan, Q. Zhang, H. Chen, J. Dong, F. Huang, and X. Huang. 2025. Next-generation database interfaces: A survey of llm-based text-to-sql. *IEEE Transactions on Knowledge and Data Engineering*.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Izacard, G., M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. 2023. Mistral 7b.

Jørgensen, T. E. and J. Breitung. 2025. Margins in contrastive learning: Evaluating multi-task retrieval for sentence embeddings. In R. Johansson and S. Stymne, editors, *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 269–278, Tallinn, Estonia, March. University of Tartu Library.

Karpukhin, V., B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Li, Z., Y. Lai, Y. Feng, and D. Zhao. 2020. Domain adaptation for semantic parsing. *ArXiv*, abs/2006.13071.

Lin, Z., Y. Yin, L. Liu, and D. Wang. 2023. Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1):315.

Litschko, R., I. Vulić, and G. Glavaš. 2022. Parameter-efficient neural reranking for cross-lingual and multilingual retrieval. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1071–1082, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Liu, M. and J. Xu. 2025. Nli4db: A systematic review of natural language interfaces for databases. *arXiv preprint arXiv:2503.02435*.

Lu, J., Y. Song, Z. Qin, H. Zhang, C. Zhang, and R. C.-W. Wong. 2025. Bridging the gap: Enabling natural language queries for nosql databases through text-to-nosql translation. *arXiv preprint arXiv:2502.11201*.

Ma, J., Z. Yan, S. Pang, Y. Zhang, and J. Shen. 2020. Mention extraction and linking for sql query generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6936–6942.

Mior, M. J. 2024. Large language models for json schema discovery.

Mohammadjafari, A., A. S. Maida, and R. Gottumukkala. 2024. From natural language to sql: Review of llm-based text-to-sql systems. *arXiv preprint arXiv:2410.01066*.

Mosca, A., F. Roda, and G. Rull. 2018. Unics-the ontology for research and inno-

vation policy making. In *Formal Ontology in Information Systems*, pages 200–207. IOS Press.

Qin, Z., Y. Song, J. Lu, Y. Song, S. Li, and C. J. Zhang. 2025. Multitend: A multilingual benchmark for natural language to nosql query translation. *arXiv preprint arXiv:2502.11022*.

Reimers, N. and I. Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Rodriguez-Penagos, C., C. Armentano-Oller, M. Villegas, M. Melero, A. Gonzalez, O. d. G. Bonet, and C. C. Pio. 2021. The catalan language club. *arXiv preprint arXiv:2112.01894*.

Sherborne, T. and M. Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153.

Shi, P., L. Song, L. Jin, H. Mi, H. Bai, J. Lin, and D. Yu. 2022a. Cross-lingual text-to-sql semantic parsing with representation mixup. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5296–5306.

Shi, P., R. Zhang, H. Bai, and J. Lin. 2022b. Xricl: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-sql semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5248–5259.

Shorten, C., C. Pierse, T. B. Smith, E. Cardenas, A. Sharma, J. Trengrove, and B. van Luijt. 2024. Structuredrag: Json response formatting with large language models. *arXiv preprint arXiv:2408.11061*.

Singh, A., M. D'Arcy, A. Cohan, D. Downey, and S. Feldman. 2023. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566.

Tola, A. 2024. *Towards user-friendly nosql: A synthetic dataset approach and large language models for natural language query translation*. Ph.D. thesis, Politecnico di Torino.

Tran, K. M. and M. Tan. 2020. Generating synthetic data for task-oriented semantic parsing with hierarchical representations. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, pages 17–21.

Wang, B., W. Yin, X. V. Lin, and C. Xiong. 2021. Learning to synthesize data for semantic parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2760–2766.

Wang, J., P. Ng, A. H. Li, J. Jiang, Z. Wang, B. Xiang, R. Nallapati, and S. Sengupta. 2022. Improving text-to-sql semantic parsing with fine-grained query understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 306–312.

Wang, L., N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. 2024. Multilingual e5 text embeddings: A technical report.

Warren, D. H. and F. C. Pereira. 1982. An efficient easily adaptable system for interpreting natural language queries. *American journal of computational linguistics*, 8(3-4):110–122.

Weller, O., M. Boratko, I. Naim, and J. Lee. 2025. On the theoretical limitations of embedding-based retrieval.

Wendler, C. et al. 2025. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.

Woods, W. 1972. The lunar sciences natural language information system. *BBN report*.

Xiong, L., C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Yang, A., B. Yu, C. Li, D. Liu, F. Huang, H. Huang, J. Jiang, J. Tu, J. Zhang, J. Zhou, et al. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Yang, J., L. Zhang, and D. Yang. 2022. Subs: Subtree substitution for compositional semantic parsing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174.

Zelle, J. M. and R. J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the national conference on artificial intelligence*, pages 1050–1055.

Zettlemoyer, L. and M. Collins. 2007. Online learning of relaxed ccg grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 678–687.

Zhang, C., Y. Li, N. Du, W. Fan, and P. S. Yu. 2019. Joint slot filling and intent detection via capsule neural networks. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5259–5267.

Zhang, X., X. Ma, P. Shi, and J. Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. *arXiv preprint arXiv:2108.08787*.

A Query Parsing Training Dataset

The synthetic data were generated from controlled vocabularies and project metadata from RIS3CAT and related R&I sources (Fuster et al., 2023). Queries were programmatically composed using domain-specific templates covering two main user intents:

- **Discover** (88.0%): exploratory searches to find relevant projects on a topic. *Examples:* "projectes sobre intel·ligència artificial", "renewable energy research in Catalonia", "proyectos de biotecnología marina".
- **Quantify** (12.0%): queries requiring counts or aggregations. *Examples:* "quants projectes H2020 hi ha sobre blockchain?", "how many universities participated in quantum computing projects?", "cuántos proyectos de economía circular en 2020?".

Resolvability and Query Types. Queries were classified by the authors according to their representability within the JSON schema:

- **Direct (77.6%)** — fully mappable to schema fields without loss of information. *Example:* "projectes H2020 de diagnòstic per imatge amb IA en col·laboració entre hospitals catalans i nord-americans des de 2019."
- **Adapted (2.2%)** — require interpretation due to ambiguous or underspecified elements (e.g., geographic scope). *Example:* "projectes de recerca a Barcelona sobre mobilitat urbana" ("Barcelona" interpreted as `province`).
- **Partial (20.2%)** — include operations that cannot be directly expressed in the JSON schema, such as ranking (*top 5 projects*) or numerical aggregation (*total funding amount*), which are partially represented in the structured form.

Distribution by Components. The component distribution was designed based on observed frequencies in real queries and the goal of covering varying levels of structural complexity. Most queries contain a thematic content element (92.8%), often combined with filters on organization type (39.9%), location (17.7%), or programme (17.6%). Temporal references (10.7%) and year-specific filters (7.8%) are less frequent.

Component Type	Frequency (%)
Thematic content	92.8
Organization type	39.9
Organization location	17.7
Programme (funding)	17.6
Time expressions	10.7
Project location	10.4
Year (specific)	7.8
Organization name	7.3

Table 10: Distribution of query components in the training dataset.

Queries vary in structural complexity, with 2–4 components being most frequent (85.1%

of all examples).

A.1 Examples of the Query Parsing Schema

Table 11 shows three annotated queries from the query parsing dataset together with excerpts of their gold structured representations. These examples illustrate the main schema elements, including filters, organisation fields, semantic queries, and intent annotations.

User Query	Gold Structured Representation (excerpt)
"ERC grants since 2020 on climate change with research centers from Catalunya"	`programme`: ERC; `year`: ¿=2020 `organisations`: [{type: research_center, location: Catalunya (region)}] `semantic_query`: "climate change mitigation and adaptation strategies" `intent`: Discover; `resolvability`: Direct
"quins projectes de la UB sobre transició ecològica?"	`organisations`: [{type: university, name: Universitat de Barcelona}] `semantic_query`: "transició ecològica" `intent`: Discover; `resolvability`: Direct
"financiación regional 2022–2024 para hospitales investigando enfermedades raras y terapias avanzadas"	`funding_level`: regional; `year`: 2022–2024 `organisations`: [{type: hospital}] `semantic_query`: "enfermedades raras y terapias avanzadas" `intent`: Discover; `resolvability`: Direct

Table 11: Examples of manually annotated queries illustrating the gold structured representation used in the dataset.

B Multilingual Query-Passage Pair Dataset

To build this dataset, we collected ˜30k scientific publications in English from several R&I and bibliographic databases, extracting their titles, abstracts, and author keywords[14]. Author keywords were originally provided as semicolon-separated lists. To obtain multilingual variants (CA, ES), we translated all fields from English using Google Translate, applied heuristic mapping, and calculate pair metadata (length ratio filtering and keyword overlap verification, and back-translation validation on a sample) to ensure consistency across languages.

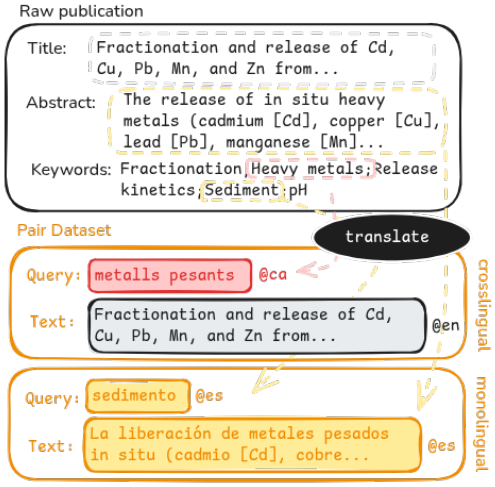

Figure 2: Examples of pairing mechanims for constructing the pair dataset.

As described in Figure 2, from these aligned documents, we constructed monolingual and cross-lingual query–passage pairs in the six directions. Queries consist of author keywords or titles, while passages correspond to titles, abstracts, or title+abstract. Most examples correspond to `keyword→text` (text as a mix of title, abstract, and title+abstract) retrieval (89.9%), while `title→abstract` pairs account for 10.1% to keep semantic similarity retrieval. Languages are balanced across splits. To enrich the corpus, each record includes linguistic metadata (token count, keyword length type, frequency quartile, and lexical vs. semantic match), enabling stratified evaluation of retrieval robustness across query characteristics such as lexical overlap, morphological complexity, and cross-lingual variation.

B.1 Examples of the Pair Dataset

Table 12 provides examples from the retrieval dataset, pairing short keyword-style queries with excerpts from the associated scientific texts.

C LoRA Training Configuration

We provide experiental detais of our baseline finetuning approaches of generative language models for query parsing. Training was run (using 1x 24 GB GPU) for all models with hyperparameter defined in Table 13 .

[14] Available at `https://huggingface.co/datasets/nicolauduran45/scidocs-keywords-exkeyliword`

Query	Text (excerpt)
Internet of Things	Internet de les coses (IoT) s'ha convertit en un pont d'informació entre societats. Les xarxes de sensors sense fil (WSN) són una de les tecnologies emergents que funcionen com a principal força en IoT. Les aplicacions basades en WSN inclouen la supervisió de l'entorn, la salut intel·ligent i la seguretat de les dades ...
redes neuronales	Síntesis de un modelo de implementación hardware de un modelo LIF de neurona biológica basado en FPGA. Tras años de investigación en neurotecnologías, se han desarrollado múltiples redes neuronales artificiales ...
Síntesi enzimàtica	Synthesis and Biological Interest of Structured Docosahexaenoic Acid-Containing Triacylglycerols and Phospholipids ...
supercomputación	Com aportar valor a les dades massives específiques del domini d'una manera interdisciplinària? Aquest treball presenta un panorama del programari i de casos d'ús d'informàtica d'alt rendiment (HPC) ...

Table 12: Examples from the retrieval dataset showing lexical (first two) and semantic (last two) query–text matches. Long texts are truncated for readability.

Parameter	Value / Description
Number of epochs	3
Effective batch size	16 (4 × gradient accumulation steps)
Learning rate	2×10^{-4}
Sequence length	2048 tokens
Precision	FP16 (mixed)
LoRA rank (r)	16
LoRA scaling factor (α)	32
LoRA dropout	0.05
Target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`
Trainable parameters	≈1% of total model weights (∼50 MB adapter)
Hardware	Single 24 GB GPU
Decoding setup	Greedy decoding, temperature = 0.1

Table 13: Training configuration and LoRA setup for model fine-tuning.

D Finetuning Hyperparameters

We provide experiental detais of our baseline finetuning approaches of sentence encoder models for content retrieval. Training was run (using 1x 24 GB GPU) for all models with hyperparameter defined in Table 14 .

Parameter	Setting
Loss function	MultipleNegativesRankingLoss
Epochs	3
Batch size	32 per device
Learning rate	2e-5 (with 10% warm-up ratio)
Selection criterion	Best checkpoint by highest R@1 on validation set

Table 14: Fine-tuning configuration for multilingual document retrieval models.

E Examples of Query Parsing Evaluation

Table 15 reports examples good, partial and incorrect parsing evaluations.

Case	User Query	Generated Model Output (excerpt)
Good	*"show me projects related to water scarcity"*	`semantic_query:` "water scarcity" `filters:` {} `meta.lang:` "en" `notes:` "No explicit filters detected" *Fully aligned with gold annotation; no spurious filters.*
Partial	*"¿cuáles son los proyectos sobre agricultura de precisión?"*	`semantic_query:` "agricultura de precisió" `filters:` {} `meta.lang:` "ca" *The semantic content and intent are correctly captured and no spurious filters are introduced. However, the generated output is generated in Catalan instead of Spanish, resulting in a language mismatch and a relaxed-but-not-strict prediction.*
Wrong	*"inversió total a Emilia-Romagna"*	`semantic_query:` "inversió total a Emilia-Romagna" `filters:` {} `meta.lang:` "ca" *The explicit location constraint is not extracted as a filter and instead goes into the semantic query, while the aggregation intent (total investment) is unsupported, resulting in a misleading structured query.*

Table 15: Examples of correct, partial, and incorrect query parsing outcomes, including excerpts of the generated structured output, generated with Salamandra 7B (base).