

Practical Prompt Engineering for AI-Supported Metadata Processing in CRIS

CRIS2026 Conference Azores – The 17th International Conference on Current Research Information Systems

Danica Zendulková

Slovak Centre of Scientific and Technical Information (CVTI
SR)
Bratislava, Slovakia

Otmane Azeroual

FernUniversität in Hagen (FeU)
Hagen, Germany

Supported by Project No. 401101DVH7 – SK4ERA II | Published in EPiC Series in Technology, 2026

The Metadata Challenge in CRIS Environments

Drivers & Pressures

- CRIS as critical infrastructure for aggregating publications, projects, and datasets.
- Open Science and FAIR data principles raise metadata quality expectations.
- Research assessment reforms demand standardized institutional data.

Current Limitations

- **Manual Curation:** Time-consuming, inconsistent, and highly error-prone.
- **Technical Gap:** High barriers for institutions lacking advanced ML capabilities.
- **Scalability:** Manual processes cannot keep pace with digitalization growth.

The Need for Low-Threshold AI

- ✓ Shift from complex ML training to accessible prompt-based solutions.
- ✓ Embedding institutional knowledge directly into AI interactions.

Slovak CRIS Case Study

Critical import bottleneck identified: research project records lacked the mandatory **"field of science and technology"** classification, blocking integration into the national system.

~1,000 RECORDS MISSING MANDATORY FIELDS

Research Objective and Paper Contributions

? Research Question

Can widely available AI tools support the assignment of appropriate scientific classifications based on project titles and descriptions in CRIS environments?

Key Paper Contributions

-  **Practical Framework:** Integrating prompt engineering into CRIS workflows without advanced technical infrastructure.
-  **Empirical Evidence:** Large Language Model (LLM) performance in metadata classification under real-world conditions.
-  **Human-in-the-Loop:** A hybrid approach enhancing metadata quality, consistency, and scalability.

>_ Concept: Prompt Engineering

Structured textual instructions guide LLM behavior by embedding institutional knowledge and classification schemes directly into interactions.

Low Threshold: No model training or heavy computational resources required.

Classification Scheme (OECD/Frascati)

LEVEL 1 Broad Domains (e.g., Natural Sciences)

LEVEL 2 Subfields (e.g., Materials Engineering)

LEVEL 3 Disciplines (e.g., Biomaterials)

Methodology and Experimental Setup

Slovak CRIS Dataset (SK CRIS)

~1,000 project records in MS Excel format. Characterized by heterogeneous structures, incomplete fields, and missing mandatory classification data.

4-Step Experimental Procedure

- 01 Reference Dataset:** Manual annotation by domain experts for ground truth taxonomy.
- 02 Tool Evaluation:** Testing ChatGPT, Gemini, and Perplexity with identical prompts.
- 03 Comparison:** Selecting the best tool based on semantic and structural consistency.
- 04 Application:** Full dataset processing followed by human-in-the-loop validation.

Prompt Design Example

```
"You are a research classification assistant. Based on the following project title and description, assign the appropriate scientific classification according to the provided taxonomy... Ensure that the classification is hierarchically consistent."
```

Framework Components

-  **Structured Design:** Integration of classification rules and controlled vocabularies.
-  **Constrained Output:** Generation aligned strictly with 3-level hierarchical taxonomies.
-  **Human-in-the-Loop:** Expert validation for semantic and structural correctness.

AI-Assisted Metadata Classification Workflow

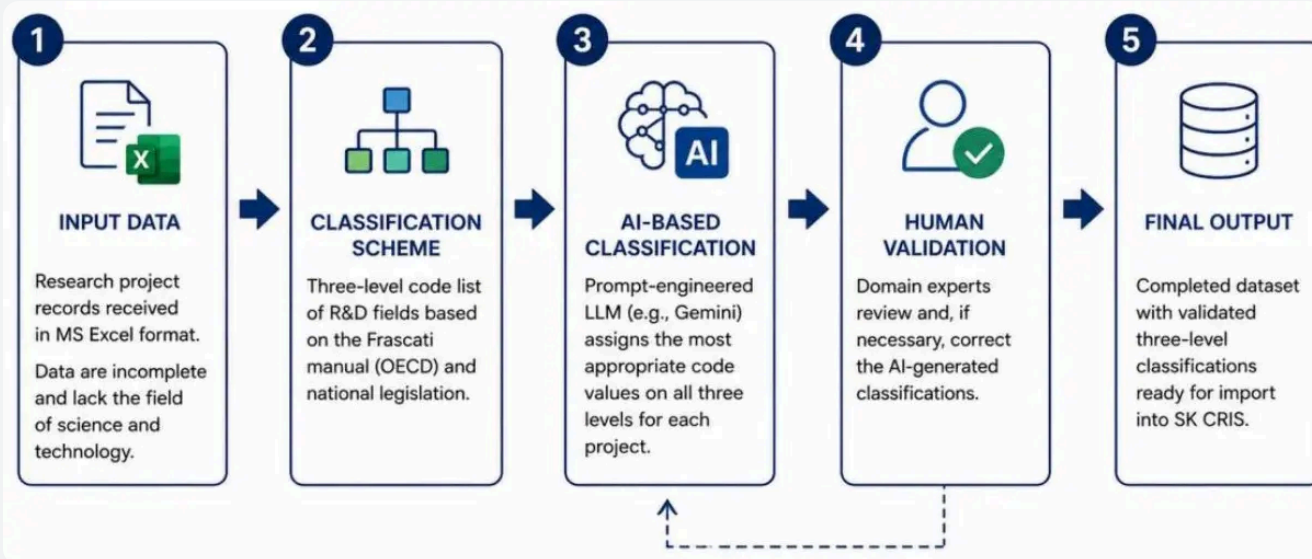


 **Figure 1: Workflow of AI-assisted metadata classification with human-in-the-loop validation**

- 1 Data Preparation**
Excel input processing & missing field identification.
- 2 Prompt-Based Processing**
Embedding classification schemes into structured prompts.
- 3 AI-Assisted Classification**
LLM generation of 3-level taxonomy outputs.
- 4 Human Validation**
Expert review for semantic and structural consistency.
- 5 CRIS Integration**
Import of validated metadata into the SK CRIS.

Key Principle: Iterative human-in-the-loop improvement ensures metadata quality and institutional alignment.

Tool Evaluation and Selection: Why Gemini?

● Perplexity **EXCLUDED**

Failed structured instructions; inconsistent outputs across taxonomy levels and incorrect cross-level mappings.

● ChatGPT **EVALUATED**

Strong semantic domain inference, but exhibited formal inconsistencies in precise institutional label naming.

✓ Gemini **SELECTED**

Superior performance with integrated taxonomy; aligned outputs with hierarchical structure and official terminology.

💡 Explicit integration of the official code list into prompt design proved decisive for constrained classification.

TASK ASSIGNMENT FOR GEMINI	
 INSTRUCTION	I will provide you with a code list of science and technology fields used in the Slovak CRIS system. Based on the project title and its description/annotation, assign the most appropriate categories from this code list. For each record, provide all three levels of the classification (Level 1, Level 2, Level 3). Use only the codes and names from the provided code list.
 CODE LIST	< Three-level code list of R&D fields (provided in the document) >
 INPUT DATA (Project data)	< Table with project title and description/annotation (provided in the document) >
 REQUIRED OUTPUT	Return the results in a table with the following columns: Project Title Level 1 Level 2 Level 3 Fill in the three levels of the classification for each project.

Figure 2: Task assignment for Gemini using structured prompt instructions and classification scheme

Results: AI-Generated Classification Outputs

PROJECT TOPIC	LEVEL 1 (DOMAIN)	LEVEL 2 (SUBFIELD)	LEVEL 3 (DISCIPLINE)
Social withdrawal research	SOCIAL SCIENCES	Psychological sciences	Clinical psychology
Molecular materials engineering	ENGINEERING & TECH	Materials engineering	Composites
Bladder carcinoma genetics	MEDICAL & HEALTH	Basic medical sciences	Medical oncology
Smart IoT energy management	ENGINEERING & TECH	Electrical engineering	Automation & control
Sustainable agriculture	NATURAL SCIENCES	Agricultural sciences	Agronomy

Figure 3: Sample of AI-generated classification results with assigned three-level taxonomy codes.

High Accuracy at Level 1

Broad domains are identified with near-perfect accuracy, providing a solid foundation for top-level CRIS categorization.

Granularity Challenges

Precision slightly decreases as granularity increases (Levels 2 and 3), requiring closer expert scrutiny at the discipline level.

Efficiency Gains

The majority of AI outputs are semantically plausible and structurally consistent, leading to a low proportion of manual corrections.

Operational Insights and Key Limitations

Operational Observations

Batch Size Effect

Smaller batches provide higher stability; larger batches lead to performance degradation or "model fatigue."

Input Dependency

High accuracy requires detailed project titles and descriptions; vague inputs result in unreliable classification.

Language & Format

Slovak language robustness is high in LLMs; however, Excel files require conversion for text-based processing.

Research Limitations

Generalizability

Findings are based on a single national CRIS dataset; application to other systems may yield varying results.

Scale Constraints

The study utilized a moderate dataset size (~1,000 records); future work must test larger, more heterogeneous data.

Scenario Breadth

Future extensions are needed for multi-institutional scenarios and complex cross-border metadata governance.

“ **Quality In = Quality Out:** AI effectively amplifies metadata consistency but cannot compensate for fundamentally inadequate input data.

Discussion: AI as an Operational Component

Knowledge Translation

Prompt engineering acts as a bridge, translating complex institutional classification rules and controlled vocabularies into machine-interpretable instructions, ensuring **transparency and controllability**.

Paradigm Shift in CRIS

AI is evolving from a purely analytical tool into a core **operational component** of metadata governance, fundamentally improving consistency and workload management.

Approach	Scalable	Training Data	Technical Threshold
Manual Curation	No	None	Expert only
Rule-based Validation	Partial	None	Medium
ML Pipelines	Yes	Large datasets	High
Prompt Engineering	Yes	None needed	Low ✓

A Reusable Prompt Engineering Framework



1. PREPARATION

- Identify missing/incomplete metadata fields
- Establish expert ground truth reference sample



2. PROMPT DESIGN

- Embed classification schemes & terminology
- Define hierarchical rules & output constraints



3. PROCESSING

- Apply LLM in iterative, small batches
- Generate structured tabular classification data



4. VALIDATION

- Expert review for semantic correctness
- Targeted correction of structural edge cases



APPLICABILITY CONDITIONS

- ✓ Taxonomy defined
- ✓ Domain expert access
- ✓ Project titles/desc.
- ✓ Digital data format



KEY FRAMEWORK BENEFITS

- + No custom training
- + Scalable to datasets
- + Low technical entry
- + Reproducible results

Conclusion and Future Directions

✓ Key Takeaways - Part I

- ✓ Prompt engineering bridges institutional knowledge and AI capabilities without complex infrastructure.
- ✓ Human-in-the-loop validation is essential for a robust metadata governance framework.
- ✓ Input data quality is a critical prerequisite; AI amplifies quality but cannot fill fundamental gaps.

✓ Key Takeaways - Part II

- ✓ A hybrid human–AI approach enables scalable metadata processing for large institutional datasets.
- ✓ High-quality metadata enables knowledge integration aligned with Open Science and FAIR principles.

➡ Future Directions

- ➡ Development of refined prompt strategies and adaptive feedback mechanisms for higher granularity.
- ➡ Extension to multilingual datasets and cross-institutional CRIS interoperability scenarios.
- ➡ Integration with automated CRIS data quality pipelines and validation frameworks.



Core Message: Prompt-engineered LLMs provide a practical, low-threshold solution for CRIS metadata governance, transforming AI into an operational component of research infrastructure.

Thank You for Your Attention

Questions & Discussion Welcome

CRIS2026 Conference Azores – Practical Prompt Engineering

Danica Zendulková

Slovak Centre of Scientific and Technical Information (CVTI SR)
Lamačská 8/a, 840 05 Bratislava, Slovakia

✉ danica.zendulkova@cvtisr.sk

Otmane Azeroual

FernUniversität in Hagen (FeU)
Universitätsstraße 47, 58097 Hagen, Germany

✉ otmane.azeroual@fernuni-hagen.de



CRIS2026 Azores – The 17th International Conference on Current
Research Information Systems