Check for updates

RESEARCH ARTICLE

# Towards AI-assisted metadata generation for improved description of geospatial data

[version 1; peer review: 1 approved with reservations]

Margherita Di Leo[1], Ilyas Tiouassiouine-Maes[2], Jordi Escriu[2], Hagar Lowenthal[2], Manlio Bacco[2], Marco Minghini[2]

[1]Arcadia SIT, under contract with the European Commission, Joint Research Centre, Ispra, VA, 21027, Italy
[2]European Commission Joint Research Centre Ispra, Ispra, Lombardy, Italy

**v1** **First published:** 06 Aug 2026, **6**:277
https://doi.org/10.12688/openreseurope.24273.1
**Latest published:** 06 Aug 2026, **6**:277
https://doi.org/10.12688/openreseurope.24273.1

## Abstract

Metadata are fundamental components of spatial data infrastructures, enabling the discovery, evaluation, and reuse of datasets. Their creation and maintenance, typically performed manually, can be costly, time-consuming, and prone to inconsistencies. This work investigates the feasibility of using Generative AI (GenAI) to assist human operators working for data producers in creating structured dataset descriptions compliant with GeoDCAT-AP standards. We address the following research question: to what extent can layered prompt engineering strategies improve the quality of metadata descriptions generated through Large Language Models (LLMs), and how do individual prompt components-role definition, content rules, template structure, and few-shot examples-interact with model selection to affect structural compliance, semantic similarity, and factual reliability?

We propose a six-layer prompting framework and evaluate seven ablation strategies, each selectively disabling specific layers, using two LLMs (qwen3-32b and qwen3-coder-30b-a3b-instruct) and eight geospatial datasets. Generated descriptions are assessed using eight automated metrics spanning lexical quality, structural compliance, and semantic similarity, complemented by targeted expert evaluation of factual accuracy.

Results reveal a clear hierarchy of layer impact. The input template, used as the reference structure of the description, is the most influential component, driving both structural compliance and readability. Few-shot examples are the second most impactful layer, substantially reducing redundancy and improving semantic

**Open Peer Review**

**Approval Status** ?

| | 1 |
|---|---|
| **version 1** 06 Aug 2026 | ? view |

1. **Zachary Sherman**, Department of Geography, Virginia Tech, Blacksburg, USA

Any reports and responses or comments on the article can be found at the end of the article.

alignment. Content rules do not measurably improve surface-level output quality, but serve as critical safeguards, encouraging models to flag missing information rather than fabricate content. Expert role definition contributes the least measurable effect. Both LLMs produce semantically comparable outputs under full guidance but diverge without structural constraints. We recommend the full prompting configuration for production use and provide practical guidelines for balancing prompt complexity, output quality, and factual reliability in LLM-assisted metadata generation workflows.

**Keywords**
Generative Artificial Intelligence, Data Spaces, Geospatial, Metadata, Large Language Model, Prompt Engineering, Spatial Data Infrastructures

This article is included in the Joint Research Centre gateway.

**Corresponding author:** Margherita Di Leo (margherita.di-leo@ext.ec.europa.eu)

**Author roles: Di Leo M**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation; **Tiouassiouine-Maes I**: Conceptualization, Data Curation, Methodology, Writing – Original Draft Preparation; **Escriu J**: Conceptualization, Writing – Original Draft Preparation; **Lowenthal H**: Data Curation, Writing – Original Draft Preparation; **Bacco M**: Conceptualization, Methodology, Writing – Original Draft Preparation; **Minghini M**: Conceptualization, Methodology, Writing – Original Draft Preparation

**Competing interests:** No competing interests were disclosed.

**Grant information:** Co-authors are member of staff at the Joint Research Centre of the European Commission.
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2026 Di Leo M *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License , which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Di Leo M, Tiouassiouine-Maes I, Escriu J *et al.* **Towards AI-assisted metadata generation for improved description of geospatial data [version 1; peer review: 1 approved with reservations]** Open Research Europe 2026, **6**:277 https://doi.org/10.12688/openreseurope.24273.1

**First published:** 06 Aug 2026, **6**:277 https://doi.org/10.12688/openreseurope.24273.1

## 1. Introduction

Data management, and in particular metadata management approaches, are continuously changing to meet the ever-evolving requirements of data sharing, visiting, and use in modern systems. At present, data architectures often fail to deliver the promised value when asked to keep up with the scope and the scale of different use cases.[1] To uphold such promises, novel approaches have been proposed because of their potential to provide needed capabilities such as flexibility, scalability, and interoperability, with emphasis on simpler and faster operations. In those new approaches, the potential of Artificial Intelligence (AI) in assisting and/or automating operations is of crucial interest. Metadata, literally 'data about data', are, among others, descriptions of datasets designed to facilitate data retrieval and use. Metadata are recognised as a core asset of the Findable, Accessible, Interoperable and Reusable (FAIR) principles for data management and stewardship,[2] which nowadays represent the foundation of data infrastructures worldwide. In all approaches related to metadata, a key objective is the generation of high-quality ones, i.e., structured, standardised, and machine-readable metadata that can reliably support automated data discovery, integration, and analysis. By creating or extracting high-quality metadata, loss of data value can be minimised. The importance of metadata is not limited to the technical realm but is also well-understood at the legislative level worldwide. In this paper we deliberately focus on the European Union (EU), which, through a phenomenon known as the Brussels effect, has acquired a unique capacity to influence international policies. In the EU, the European Commission included specific provisions on metadata in legal acts governing data sharing and reuse such as the INSPIRE[3] and Open Data[4] Directives, and the Data Governance Act.[5]

From an economic viewpoint, the loss of data value because of poor metadata quality is a threat to the establishment of a lively data market.[6,7,8] In the same context, which has seen massive investments of the EU in the creation of a single, unified data market to foster a data-driven economy, the establishment of Common European Data Spaces (CEDS), as envisioned by the European Strategy for Data,[9] makes metadata essential to enable data discovery, sharing, reuse, and interoperability across different domains and sectors. Because of the opportunities that such a market has been providing, further investing in high-quality metadata is strongly needed to enhance discoverability and, in turn, open new opportunities for improved decision making, efficiency gains, and ultimately societal innovation.

However, generating and maintaining high-quality metadata has a cost to be carefully factored in the cost of data provisioning,[10] since it is typically a resource-intensive and time-consuming process. When effort is minimised in carrying out this task, this often results in a lack of completeness, consistency, accuracy, and updates. While manual curation typically yields optimal results, such an approach is not always viable, particularly when dealing with large volumes of similar datasets that require batch metadata creation, which makes the process time-consuming and error prone. Even when significant effort is put in provisioning high-quality metadata, some issues are likely to persist. This, to some extent, is an inevitable consequence of the fact that metadata are context-specific and depend on the sectorial domain and the audience for which they are provided.[11] Taking the case of CEDS, and considering data sharing among different sectorial data spaces, additional challenges are foreseeable considering that data is collected subject to possibly diverging frameworks and standards (e.g., because of the different focus of the data spaces) and that metadata of datasets in a data space should be usable also in other data spaces. In addition, the growing set of multimodal – i.e., in different formats – data requires the use of multimodal metadata understanding techniques.[12] In the end, without proper metadata management able to correctly handle interdependencies, relationships, and semantic structures, the quality of metadata may severely degrade,[12] hindering the potential of leveraging data in an effective way in the data market.

The use of AI may offer a promising solution to some of these challenges, with the potential to facilitate and even automate various aspects of the metadata lifecycle, from creation and curation to validation and exploitation. In this work, we explore the potential of Generative AI (GenAI), specifically Large Language Models (LLMs), to support users in metadata extraction and generation, proposing a human-in-the-loop approach. The objective is to get metadata generation closer to be AI-ready. When it comes to metadata generation, AI readiness could be translated as the possibility to rely on metadata standards, which include schemas and vocabularies, to generate high-quality and domain-specific metadata in a consistent manner using GenAI, thus with minimal human intervention.

To test the idea above, we consider the specific case of geospatial data and corresponding metadata. We do this in light of our longstanding experience in the management of such data, through the work of the Joint Research Centre (JRC) as the coordinator of the technical implementation of the EU INSPIRE Directive,[13] where high-quality metadata are key to ensure proper data discoverability and accessibility.

The remainder of the paper is structured as follows. Section 2 reviews existing work on the use of AI for metadata creation and discusses the limitations of current metadata standards for direct use in LLM-assisted generation. Section 3 introduces the metadata lifecycle and identifies the main challenges in metadata creation that motivate AI-assisted approaches. Section 4 presents our methodology, including how metadata attributes are classified to guide the generation

process, the pipeline used to produce dataset descriptions, and the prompting framework designed to systematically test how different prompt components affect output quality. In section 5, the evaluation metrics are defined. The results are presented in Section 6, where the impact of individual prompting layers and model differences are analysed. Section 7 discusses the implications of our findings for semi-automated metadata workflows, addresses the identified metadata challenges, and acknowledges the limitations of the study.

## 2. Related work

Several studies deal with the use of AI techniques in metadata management. In particular, Yang et al.[12] describe the metadata lifecycle that we also consider and refine in this work (see Section 3), focusing on four phases: acquisition, cleaning, verification, maintenance. Activities in those phases are crucial for metadata to be of high-quality, thus valuable to users and a data market. The authors propose a set of solutions to support, and even automate, the different activities in each of these phases, evaluating the potential of tools available in the market or developed by the open-source community.

Limiting the scope to scientific works employing the use of AI techniques, especially GenAI, for the creation and extraction of metadata from data products, a few works can be identified such as,[14–17] which mainly focus on the case of digitising libraries and how AI can be used to support such a task. Use cases of interest cover the indexing of digital resources and subjects classification, especially when it comes to the large throve of research data. For instance, Zhang et al.[17] focus on metadata annotation considering records from Research Data Australia and propose a set of prompts to classify them into categories. The authors propose a valuable discussion of three research fields in which Large Language Models (LLMs) are used to classify metadata: automated classification, use of LLMs as annotators, and in-context learning. Automated classification refers to techniques used, for instance, in expanding the accessibility of scientific knowledge, such as categories used by scientific journals to classify accepted papers. The use of LLMs as annotators refers to their use in Machine Learning (ML) tasks, such as tweet annotation for subsequent analysis. The authors conclude by saying that biases are likely in such a process because of the inability of LLMs to discern nuances and context-depending meanings. Finally, in-context learning refers to the sole use of textual interactions with a model aimed at improving results in Natural Language Processing (NLP) tasks. In other words, in-context learning can be described as the use of prompts to interact with a model and improve results.[17]

Other works highly relevant in this context are the ones in,[16,18,19,20] especially from a technical viewpoint. Martorana et al.[16] introduce a framework that automatically generates descriptions of tabular data, adopting a data-driven approach that summarises dataset contents and leverages LLMs to enrich summaries with semantic information and produce comprehensive, accurate, readable, and concise human-readable text. A controlled vocabulary is used to limit the scope of LLMs in suggesting topics for classification, a technique used in the literature to support the FAIRification of data. The authors state that contextual information (i.e., additional information about a dataset) does not make much difference in the classification results, as opposed to our experience, where providing focused and relevant contextual information contributes to more accurate outputs. Finally, in,[21] Zavalin & Zavalina consider the use of three different LLMs for metadata creation in the context of educational websites. They conclude that the quality of AI-generated results, although in the favourable context of metadata creation for websites, does not meet expected quality standards, thus still requiring the involvement of humans. Zhang et al.[18] extend the work of Martorana et al.[16] focusing on metadata creation, particularly enrichment, starting from tabular data and using different LLMs. In,[19] Nair et al. make use of LLMs to generate metadata of copies of online content in web archives. Similarly to what we propose in this work, they design a prompt to generate a structured output, providing clear instructions to the LLM thus improving the generation of metadata so that they are pertinent, concise, and precise. In,[20] Lu et al. focus on the need to search for datasets in different data provider platforms for the purpose of ecological research. To identify useful datasets, researchers must browse several sources, very often dealing with different metadata standards and availability. To ease such a work, the authors propose a harmonisation strategy, which generates a set of metadata fields according to a schema they designed, making use of an LLM-powered harvester. Both the cases of structured and unstructured metadata of online resources are considered, with the latter scoring lower than the former in tests because the lack of structure provides less context to the LLM in use.

A closely related topic is the suitability of metadata standards for direct use in GenAI-assisted metadata generation. Despite the widespread adoption of standards and the extensive documentation accompanying them,[22] existing standard documentation is poorly suited for direct use in metadata generation assisted by Language Models (LMs).[23,24] This limitation arises from two main factors: many schema standards specify attributes as free-text values, and definitions are often self-referential or provide insufficient contextual meaning for GenAI to process. Attributes used with free-text values offer flexibility, but at the cost of vagueness, further increasing the risk of errors and inconsistencies in metadata. When it comes to vocabulary definitions of metadata attributes, those are often circular or contextually shallow, providing little meaningful semantic guidance for either human metadata producers or GenAI tasked with generating or refining metadata. Such definitions, while supporting a wider adoption of metadata standards, when combined, prevent GenAI

from relying on the documentation as a consistent and trustworthy source for metadata generation. In fact, GenAI-based tools usually return vague and inconsistent outputs (low-quality metadata) when confronted with free-text attributes, unless strategies are used to guide such a process, as we do in this work. To conclude, we also highlight that the heterogeneity in the implementation of standards across domains leads to changing definitions based on different contexts,[25] further reducing both efficiency and reliability of GenAI-based approaches. What above suggests that the effectiveness of GenAI in metadata generation depends not only on its capabilities, but also on the structure, clarity and contextual depth of the metadata standards.

Based on such premises, we focus our work on the use of GenAI to support the FAIRification of data, considering the case of geospatial information, which is not covered in the literature to the best of our knowledge. Fully aware of the importance of generating high-quality metadata, we propose a human-in-the-loop approach in order to meet the needed standard for high-quality metadata while minimising the human effort required in such a task thanks to the use of GenAI techniques.

## 3. Metadata lifecycle and related challenges

In this section, we first outline the metadata lifecycle as a framework for situating the different stages involved in metadata management. We then identify the main challenges encountered during metadata creation, which provide the motivation for the AI-assisted approach presented in Section 4.

### 3.1 Metadata lifecycle

The metadata lifecycle provides a structured framework for managing the creation, dissemination, and utilization of metadata within data sharing ecosystems, particularly in the context of geospatial data. A metadata lifecycle focused on the different stages involved in the generation of metadata is proposed in[12]: acquisition, cleaning, verification, and maintenance. Building on this work, the experience gained through the management of geospatial metadata in the scope of the INSPIRE Infrastructure,[13] and the evolution towards adopting modern data ecosystems such as data spaces, we propose a wider metadata lifecycle which encompasses three core phases: metadata generation, exposure, and exploitation (see Figure 1). These phases should collectively ensure that metadata remain accurate-faithful to the data assets they describe; accessible-through a catalogue system; and actionable-capable to provide access to the data. It is generic by design to be adaptable to diverse data ecosystems, such as digital libraries, Spatial Data Infrastructures (SDIs) and data spaces, including interoperable processes aligned with the relevant standards.

The metadata generation phase comprises tasks aimed at producing metadata records and getting them ready for their publication, beginning with their first *creation* and their maintenance over time (*edition*), often carried out by the same data producers. These steps involve documenting critical attributes of data assets, e.g. key identification aspects, content

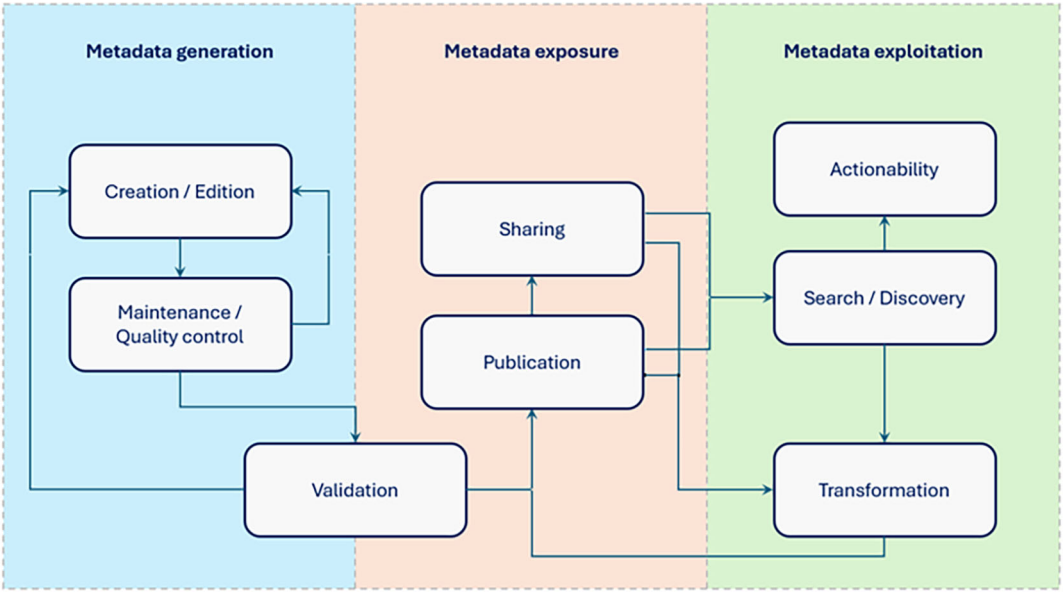



**Figure 1. The metadata lifecycle: generation, exposure, exploitation, and related inner operations. Source: own elaboration.**

description, technical specifications (e.g. coordinate reference systems for geospatial data, representation model and format), quality assessments, access conditions and use constraints, distribution channels, and links to supplementary documentation. Producers must adhere to agreed metadata standards – common frameworks for describing and structuring metadata, facilitating the discovery, sharing, and integration of data from diverse sources, such as ISO 19115[26] and GeoDCAT-AP[27] for geospatial metadata – while incorporating ecosystem-specific requirements. Subsequent steps in the *generation* phase-*maintenance*, *quality control*, and *validation*-ensure metadata remain up-to-date and compliant with evolving standards, requirements and data changes.

Following *generation*, the *exposure* phase focuses on making metadata accessible to users via *publication* and *sharing* in metadata catalogues, portals, and data discovery APIs. Publication here translates into the first public exposure of metadata records, usually by the data producer. Sharing relies on specific mechanisms to implement the reuse principle from FAIR, such as tiered harvesting or cross-ecosystem synchronization, which further enhance data visibility and impact by propagating metadata across regional, national, and international platforms. Finally, the *exploitation* phase enables users to discover, interact with, and transform metadata records. The latter option allows converting metadata into alternative formats, by third parties different from the (meta) data producer, to align with specific user needs or software requirements. Discovery tools and actionability features-such as APIs for accessing geospatial datasets-facilitate data reuse while adhering to the FAIR principles.[2]

While the overarching goal is to enhance metadata's usefulness through automation, standardization, and user-centric design, the scope of this paper is limited to exploring the use of GenAI in metadata *generation.*

## 3.2 Challenges in metadata creation

Metadata creation techniques have traditionally relied on manual annotation, which is prone to inconsistencies, highly dependent on different producers' perspectives and operators' expertise. This approach is costly, requiring significant human effort and specialized knowledge, often poorly incentivised, leading to heterogeneous metadata quality levels,[28] including accuracy. In contrast, assisted metadata generation offers a more efficient and scalable solution, leveraging rule-based systems and ML pipelines. While early approaches had limitations in terms of accuracy, recent advances in NLP, embedding techniques,[29] and LLMs (as investigated in this paper) have the potential to produce significantly enhanced, fluent and readable metadata descriptions, leveraging broad world knowledge.[18]

Ideally, metadata should support evaluating whether data is fit for use and purpose. However, the use and purpose of data can change over time and with context. Increasingly, the aim for which data was originally collected is transcended and the corresponding metadata may result inappropriate for new use cases. The fitness for use of data is likewise an evolving variable. Technological transformations imply that the use intended at the time of publication may become inappropriate at some point. Thus, ensuring that data is fit for use is a task dependent on the maintenance of relevant technological aspects. Nonetheless, to maximise re-use, relevant information in this respect should be recorded as metadata. Understandably, these are complex aspects that this paper does not consider. Instead, below we list a few common, well-known challenges in metadata creation the use of AI can help overcome, as further demonstrated in Sections 4 and 5.

The challenges can be grouped into three categories: those inherent to the metadata schema used (in Sub-section 3.2.1), those focused on metadata content (in Sub-section 3.2.2), and some more general ones (in Sub-section 3.2.3) related to the generation phase of the metadata lifecycle.

### 3.2.1 Metadata schema

- **Metadata conformity and completeness** – Due to diverse and inconsistent existing methods for editing metadata records, these may not comply to the rules defined for a given metadata schema (e.g. incorrect encoding for fields, invalid field types, order of attributes in the metadata record, inappropriate use of relationships, and presence of mandatory attributes). Furthermore, metadata schemas normally either leave the choice of using attributes as optional or only impose as mandatory the most fundamental ones. In turn, attributes that are not defined as mandatory may be missing, thus decreasing the completeness of the information provided. Such issues may also go beyond schema issues, when the metadata shall additionally adhere to specific conventions (e.g. metadata specifications and/or ad-hoc rules defined in a metadata profile). The automation offered by AI algorithms can become a useful solution not only to avoid schema conformity issues from the beginning, but also to overcome them in non-compliant records as part of validation tasks.

- **Semantics** – When generating metadata following a well-defined schema, it is not uncommon for attributes to be "misused" either by incorrectly inserting values (e.g. not using mandatory taxonomies or code lists) or by

misinterpreting their intended use. This can happen for several reasons, including ambiguous or sui-generis definitions of the schema attributes and how they should be used, as well as schemas not offering matching attributes to encode provided information. When the metadata is offered following one schema and then transformed into another, these issues are inevitably maintained and potentially amplified.

- **Use of free-text (literals values)** – It is common to proceed on a "best-intention" basis, when generating textual values such as descriptions, abstracts or titles. Although supported by a metadata schema, validating their quality or correctness can be challenging. The texts provided are often too brief, too long or specialised, or context specific. AI NLP and LLM technologies can assist automating these processes.

3.2.2 Metadata content

- **Blank-sheet syndrome** – At the time of generating new metadata content, it is common to face difficulties to start the edits, due to the need to find relevant information from existing sources and to reflect on the best way to describe the datasets. This situation tends to diminish edition efficiency at the beginning of the process, progressively improving while the task advances. AI may offer metadata editors tools to bypass this syndrome, offering initial metadata field contents to be refined and improved through human knowledgeable intervention.

- **Logical (internal) consistency** – Content included in a single metadata record may suffer from internal consistency issues due to lack of knowledge and/or deficiencies in the metadata generation phase (e.g. incongruent or conflicting statements, inconsistent values for creation, edition and publication dates). Detecting such problems might be complex, ineffective and/or inefficient through manual quality controls, where AI tools may play a helpful role instead.

- **Consistency between related metadata records** – When a set of datasets are sharing common attributes or belonging to the same data planning strategy, their metadata records shall keep specific consistency rules to make this set meaningful for users. Separately editing these records often tends to overlook such rules and makes it difficult to comply homogeneously with them for all records.

3.2.3 General

- **Data quality** – Very often, quality assurance of the data is not explicitly encoded in the metadata. Even when it is part of the data specifications and/or the methodological documentation of data production, this information is commonly overlooked in the metadata records. AI algorithms may help document data quality based on existing quality checking reports and relevant existing documentation and mapping the results to defined attributes of specific metadata models.

- **Thematic classification** – This operation tends to be done inconsistently across domains, sectors and application fields. The task requires substantial effort and, ideally, would require a systematic approach. The use of AI could, for example, enrich metadata by using equivalent terms from distinct vocabularies, or enrich free-text terms used in the metadata with similar ones. Furthermore, adopting AI for this task would imply using a systematic and consistent method for all instances being classified.

**Table 1. Classification of GeoDCAT-AP metadata attributes.**

| Type | Name | Description | Examples | Generation |
|---|---|---|---|---|
| A | Descriptive | Free-text fields conveying semantic content | Title, description, keywords, abstract | LLM-assisted generation |
| B | Technical | Data-intrinsic properties extractable from the source | Geospatial extent, Coordinate Reference System, spatial resolution, file format, creation date | Extraction using specialised libraries |
| C | Institutional | Properties defined by the data provider | Access rights, license, distribution, contact point | Defined by the organisation, often reused across datasets |

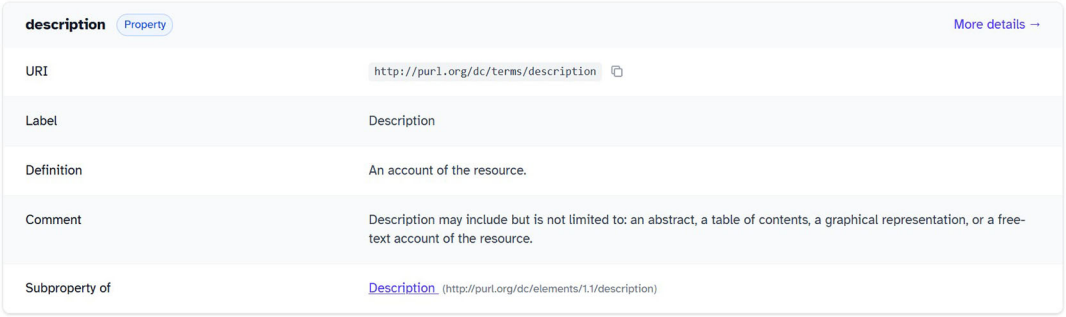



**Figure 2. Definition of the metadata attribute "description" for the Dataset entity as in the Dublin Core vocabulary[31] and referenced in the GeoDCAT-AP documentation.[27]** Source: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#description

- **Metadata in data** - Some elements that may be considered metadata are an integral part of the data contents such as, for example, column headers of tabular data or coordinates of tiles. Conversely, it may be useful to iterate some elements encoded as part of the data contents in the metadata. For example, data measuring terrain elevation should be accompanied with metadata referring to the geospatial area the measurements refer to, to help users quickly assess its relevance. Cases like these could benefit from automated knowledge extraction provided by GenAI.

## 4. Methodology

As mentioned in Section 3, in this work we explore the application of GenAI for the generation of the content of metadata attributes. More specifically, we analyse how prompt engineering strategies and model selection affect the quality of automatically generated metadata descriptions.

### 4.1 Attributes taxonomy

We focus on the GeoDCAT-AP v3.1.0 metadata schema,[27] an extension of the DCAT vocabulary[30] for geospatial data. Based on their nature, the attributes of the GeoDCAT-AP schema can be classified into three types (see Table 1 and Extended data section).

- Descriptive attributes (Type A) are free-text fields suitable for NLP and LLM processing, such as title, description, and keywords.
- Technical attributes (Type B) can be extracted directly from the data source and include geospatial extent, Coordinate Reference System (CRS), spatial resolution, file format and creation date.
- Institutional attributes (Type C) are determined by the data provider and are typically shared across multiple datasets, and include producer name, contact point, access rights and license.

We focus our attention on the "description" attribute, a mandatory free-text field providing a human-readable account of the geospatial dataset. Among Type A attributes, it is the most information-rich and also the most dependent on Types B and C attributes for its content, making it the most demanding attribute to generate and therefore a meaningful target for assessing the effectiveness of prompt engineering strategies.

### 4.2 The generation pipeline

The "description"[1] attribute of the GeoDCAT-AP schema reuses the property of Dublin Core,[31] one of the most widely used web ontologies[2]. As shown in Figure 2, the definition and explanation of the "description" attribute of the GeoDCAT-AP schema (inherited from Dublin Core) are rather vague, which in turn provide a high degree of freedom to the metadata editors having to fill its content.

[1]https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#description
[2]See: https://lov.linkeddata.es/dataset/lov, which shows the Dublin Core and Dublin Core Terms ontologies as the most reused ontologies (481 and 782 outgoing links respectively).

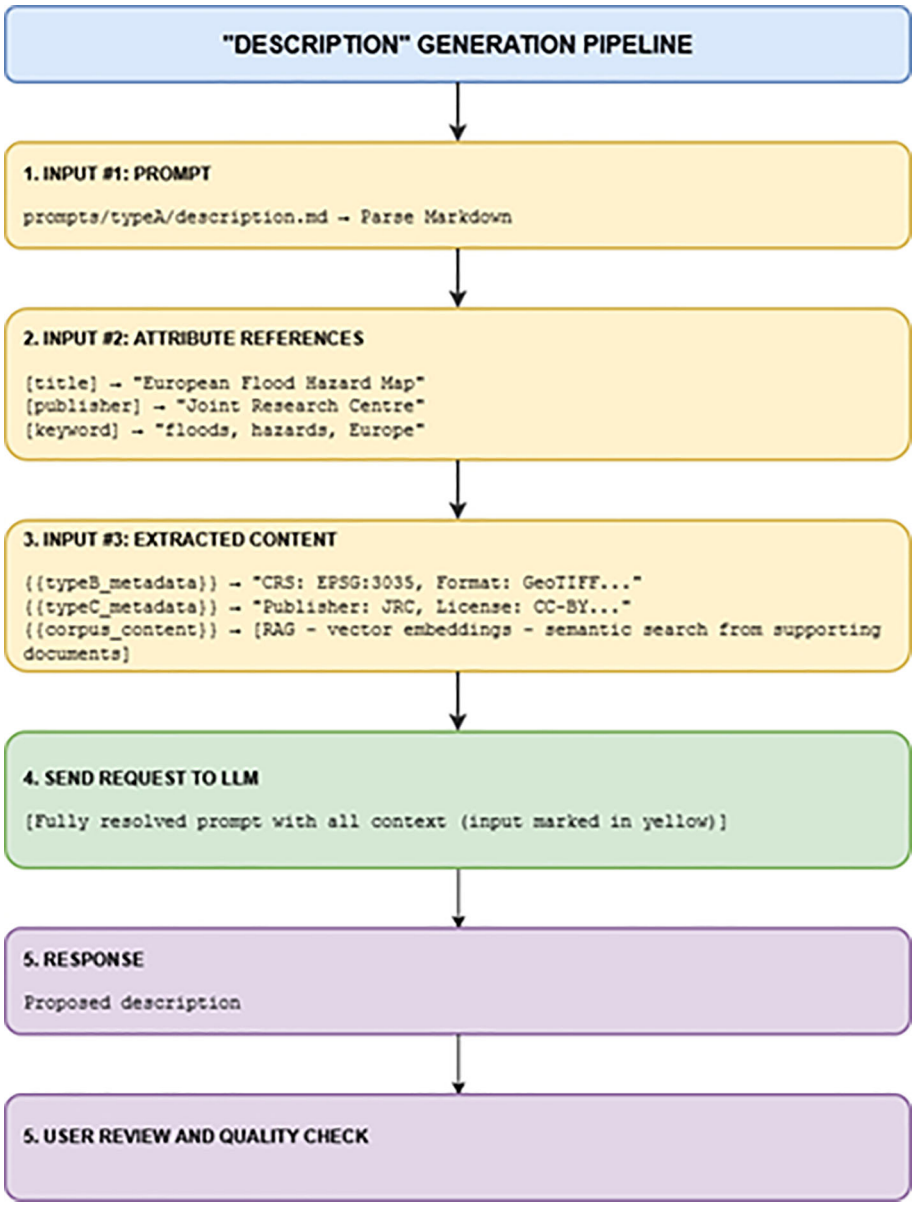



**Figure 3. Pipeline for the generation of the description attribute.** Source: own elaboration.

In our approach (depicted in Figure 3), the description is generated using (1) a prompt and additional contextual information, which is: (2) attribute references (e.g., title, keywords, publisher, etc.); (3) content extracted from the dataset (Type B attributes such as Coordinate Reference System, dataset format, spatial resolution), content extracted from a publisher document (Type C attributes such as publisher, licensing), and a corpus generated from supporting documents related to the dataset, such as technical reports or scientific publications. The prompt and contextual information are sent to the LLM through an API request. The proposed description is shown to the user (human in the loop) for a quality check and approval.

### 4.3 Strategy combinations for testing

Before describing the strategy combinations, we clarify key terminology used throughout the remainder of the paper. Our approach relies on a structured template containing a set of **expected fields** (e.g., [*geographicalCoverage*], [*temporalCoverage*]) that the LLM is instructed to populate with content extracted from the source material. When the required information is not available in the sources, the LLM is expected to replace the field with the marker [*Not specified*]. We refer to such markers in the generated output as **placeholders**, indicating willingly unfilled information gaps. The

Placeholder Ratio, introduced in Section 5, accordingly measures the proportion of expected fields that remain as placeholders in the output: a lower value indicates more complete content, while a higher value indicates greater transparency about missing information.

4.3.1 Prompting Strategy Layers

To generate a suitable prompt, we designed a prompting framework consisting of six configurable layers (see Figure 4). Each layer can be enabled or disabled, and we describe those in what follows.

- **Layer 1 - Role Definition** - establishes the persona and behavioural constraints for the LLM. We considered the following two options: 1) Expert: "You are a geospatial metadata expert specialized in producing GeoDCAT-AP compliant dataset descriptions". 2) None: No role definition provided.

- **Layer 2 - Task Specification** - consists of instructions on what should be generated – It is the same across all variants, in the form: "Generate a dataset metadata description compliant with GeoDCAT-AP."

- **Layer 3 - Content Rules** - Anti-hallucination constraints specifying what should be included. Options: 1) Enabled: "Extract ONLY information explicitly stated in sources. DO NOT infer, generalise, or assume. If information is missing, write [Not specified]." 2) Disabled: No explicit content constraint.

- **Layer 4 - Context Injection** - Provides source material for generation. It is the same across all variants. Content: Type B attributes (technical information), Type C attributes (publisher information), Corpus content (supporting documents).

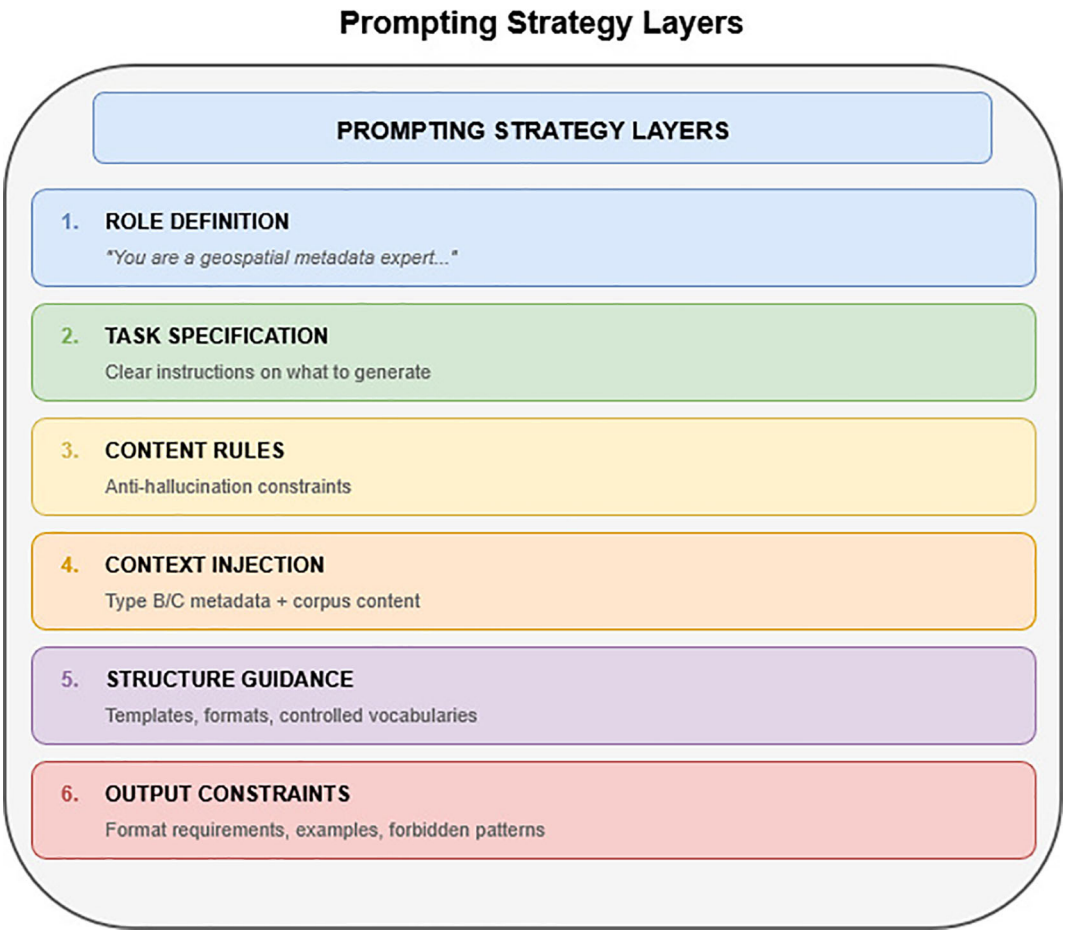



**Figure 4. Prompting strategy layers considered in this study.** Source: own elaboration.

- **Layer 5 - Structure Guidance** - Template, format, and structural requirements of the output. Options: 1) Template and Rules: rigid template provided with detailed placeholders (rules accompanied by examples of which values the placeholders can contain). 2) Template only: rigid template without detailed extraction rules. 3) Free form: no template, i.e., free text generation.

- **Layer 6 - Output Constraints** - Format requirements, examples, and forbidden patterns. Options: 1) Full (Positive and Negative): positive examples (what can be done) and negative examples (what cannot be done). 2) Partial: only positive examples. 3) None: no examples.

It is to be noted that each additional layer increases the length of the prompt sent to the LLM. Since LLMs have finite input capacity (measured in tokens, i.e., word fragments), overly complex prompts risk exceeding these limits or increasing processing costs. Therefore, identifying the minimal effective set of layers is not merely an academic exercise -it has practical implications for deploying LLM-assisted metadata generation at scale, balancing output quality against computational efficiency.

4.3.2 Strategy configuration matrix

Following the pipeline described in Figure 3 and then using the different prompting layers described in Figure 4, we are now ready to test different combinations and analyse the results. Seven strategy combinations (see the columns of Figure 5) have been designed to systematically test the impact of different prompting layers with controlled variation by a layered ablation design.

4.3.3 Strategy descriptions

The seven strategy combinations are summarised in Figure 5. **S1** (**Full Expert**) serves as the reference standard, with all six layers enabled. The remaining strategies each disable specific layers to isolate their contribution. **S2** (**No role**) removes the expert role definition while keeping all other layers identical to S1. **S3** (**Expert Minimal**) retains the template structure but removes both extraction rules and few-shot examples, testing whether the template alone is sufficient. **S4** (**Expert No Rules**) removes the anti-hallucination content rules, testing whether their absence leads to more complete but potentially less reliable outputs. **S5** (**Expert Free Form**) removes the template entirely, allowing free-form generation under expert role and content rules. **S6** (**Expert Pos Only**) drops only the negative examples, testing whether positive examples alone provide sufficient guidance. Finally, **S7** (**Baseline**) retains only the task specification and context injection, establishing a minimal baseline against which all other strategies are compared.

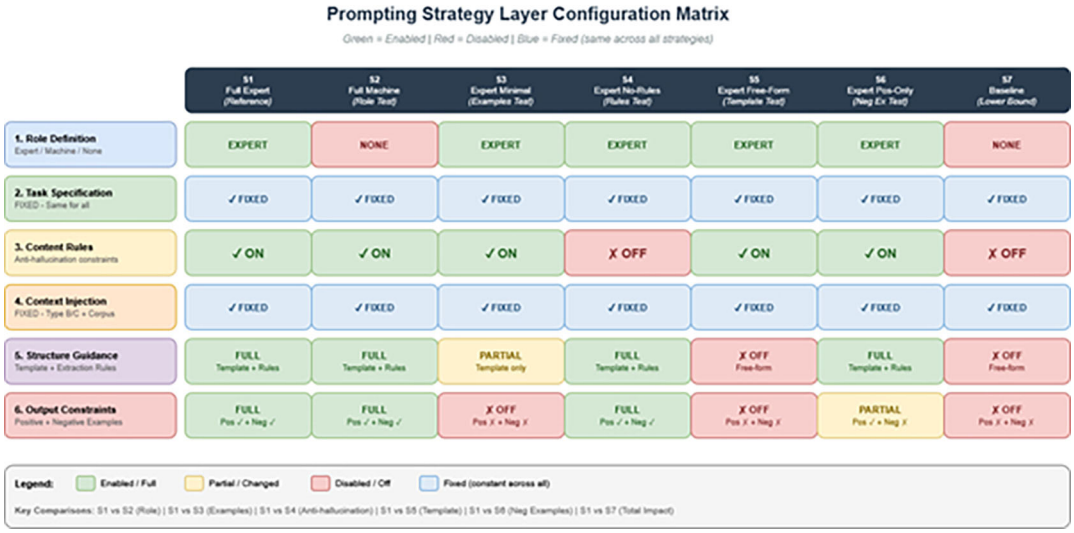



**Figure 5. Configuration matrix of the prompting strategies.** The layers in Figure 4 are visible on the rows, while the strategies (S1–S7) are on the columns.

## 4.4 Experimental setup

The experimental setup also considers two LLMs, specifically a general-purpose model such as qwen3-32b (Model1); and a code-specialised model, such as qwen3-coder-30b-a3b-instruct (Model2).

The effects of the seven prompting strategies and the two models have been applied over 8 datasets for which descriptions created by human experts are available (we refer to those as "gold standard descriptions", consistently with the literature where human-authored references serve as the benchmark for evaluating generated text quality[34]).

### 4.4.1 Research questions

With this setup, we address the following questions, derived from the overarching research question formulated above:

**RQ1 (Prompt strategy effect)**

Which combination of prompting layers (role definition, content rules, template structure, few-shot examples) produces the highest-quality descriptions in terms of factual accuracy, semantic similarity, and structural compliance?

**RQ2 (Layer contribution)**

What is the relative contribution of each individual prompting layer to output quality (ablation)?

**RQ3 (Model effect)**

Does a general-purpose LLM (qwen3-32b) outperform a code-specialised model (qwen3-coder-30b-a3b-instruct), or vice versa, for metadata description generation, and how the model choice interacts with the prompting strategy?

### 4.4.2 Prompt template

The prompt template was designed to define *a priori* the expected structure and content of the "description" attribute. To create this template, we analysed existing GeoDCAT-AP descriptions from the JRC Data Catalogue,[32] which provides access to over 3,800 datasets produced by the JRC.

We selected datasets whose descriptions were associated with peer-reviewed publications containing abstracts, as our method requires documentary sources from which to extract information. This selection criterion ensured the availability of rich, structured textual content to inform the template design.

Using GPT-4o,[33] we identified recurrent structural patterns across the selected descriptions-such as the presence of geographic scope, temporal coverage, methodology, and intended use. These patterns were then manually consolidated into a prompt template containing placeholders to be populated by the LLM using the target dataset and its supporting documentation. This template is referred to as *[template]* in the following sections.

It should be noted that the template design and the generation experiments are independent tasks: the template defines *what* information a description should contain, while the experiments evaluate *how effectively* different models and prompting strategies populate it. Using a different model for template design than for generation avoids circularity and ensures that the template is not inadvertently optimised for a specific model's behaviour.

```
"This dataset provides [broad topic of dataset content]. It depicts [topic or
theme] for [geographicalCoverage] for [temporalCoverage], with a resolution of
[spatialResolutionAsText].
The data have been produced using [methodology] and sourced from [data sources].
The data are available in [format] and have [dataset representation].
Potential applications include [intended use or application].
The dataset can be accessed at [landingPage], with the following access rights:
[accessRights]. For enquiries, contact [contactPoint]. The dataset is updated
[frequency].
NOTE: [limitations or notes if any]"
## How to extract placeholders
### Auto-filled from metadata (use as provided, or "[Not specified]" if empty)
```

```
-[geographicalCoverage] - Spatial coverage from extracted metadata
-[temporalCoverage] - Temporal period from extracted metadata
-[spatialResolutionAsText] - Resolution description from extracted metadata
-[format] - File format from Type B metadata
-[landingPage] - Access URL from Type C metadata
-[accessRights] - Access conditions from Type C metadata
-[contactPoint] - Contact information from Type C metadata
-[frequency] - Update frequency from Type C metadata
```

**Template 1** Generated template to populate the content of the "description" attribute of the GeoDCAT-AP schema. It contains 8 expected fields.

## 5. Evaluation

The 112 generated descriptions - seven prompting strategies applied across two models and eight datasets - have been evaluated through a set of automatically extracted 8 quality metrics (BERTScore provides 3 metrics), divided in 3 categories, as shown in Figure 6.

### 5.1 Lexical metrics

Lexical metrics measure surface-level text quality characteristics including readability, vocabulary diversity, and repetition patterns.

- The Flesch-Kincaid Grade Level[35] evaluates text complexity based on sentence length and word complexity (syllable count). Scores between 10–14 are considered appropriate for technical/scientific content.
- The Flesch Reading Ease score[36] rates text on a 0–100 scale, where higher scores indicate easier readability. It uses the same core measurements as Flesch-Kincaid (sentence length and syllables per word) but with different weighting factors. Scores ranging between 30–50 are considered appropriate for technical/scientific content.
- The Type-Token Ratio[37] measures lexical diversity (vocabulary richness) by comparing the number of unique words (types) to the total number of words (tokens). While values approaching 1.0 indicate maximum vocabulary diversity, TTR decreases naturally with text length. For texts of the length typical of metadata descriptions, values between 0.5 and 0.8 indicate adequate lexical diversity without excessive repetition.
- The Redundancy Score[38] measures text repetitiveness by analysing bigram (two-word sequence) patterns. The metric calculates how many bigrams are repeated versus unique, with lower scores indicating less repetition and better text quality. (0.0–0.1 Excellent; 0.1–0.2 Good; 0.2–0.3 Moderate; >0.3 High redundancy).

### 5.2 Structural metrics

Structural metrics evaluate adherence to required templates and detect missing content as placeholders.

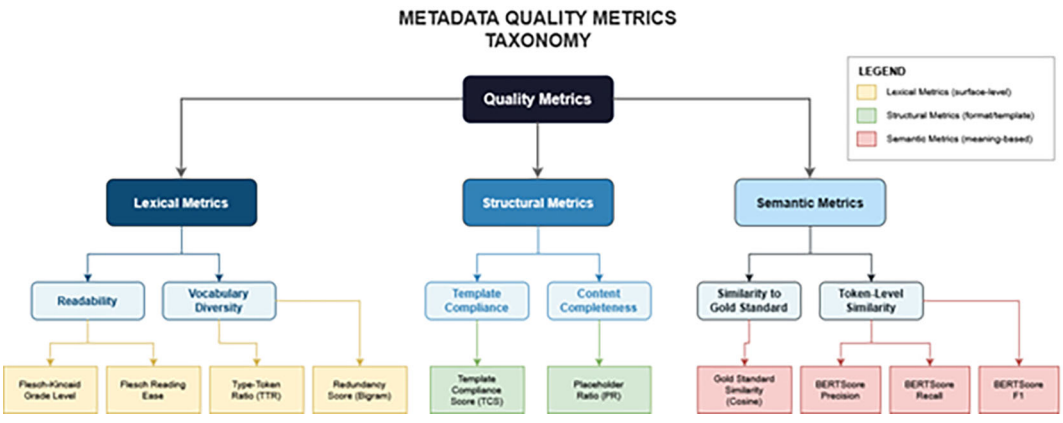



**Figure 6. Quality metrics taxonomy: lexical, structural, and semantic metrics.**

The Template Compliance Score (TCS) operationalises the metadata completeness dimension defined by Bruce and Hillmann[39] and automated by Ochoa and Duval[40] for digital repository metadata. It adapts the Decomposed Requirements Following Ratio proposed by Qin et al.[41] for LLM instruction-following evaluation, measuring the proportion of expected template components present in the generated output. It is calculated as follows:

TCS = (Number of present components)/(Total required components).

Scores vary from 1.0 (full compliance) to <0.6 (poor compliance).

The Placeholder Ratio (PR) extends the completeness assessment by detecting explicit information gap markers, such as [*not specified*] placeholders, following the automated quality assessment approach of Neumaier et al.[42] for open data portal metadata.

It is calculated as follows:

$$\mathrm{PR} = (\text{Number of placeholders found})/(\text{Number of expected fields})$$

With placeholders we refer to missing values, thus a lower value of the index translates into more complete content. A value of 0 represents the ideal case where all expected fields have been populated with extracted content. In practice, however, non-zero values are expected and not necessarily indicative of poor performance, as source documents may legitimately lack information for certain fields. In such cases, an explicit placeholder is obviously to be preferred to fabricated content, as later discussed in Section 6.3.

### 5.3 Semantic metrics

Semantic metrics evaluate meaning-based similarity between generated text and reference documents using neural network embeddings.

Gold Standard Similarity[43] measures semantic similarity between the generated text and a human-written reference (gold standard) using dense vector embeddings from a pre-trained language model. Unlike lexical metrics, it captures meaning-equivalence even when different words are used.[43,44] The embedding vectors of the generated and reference texts are computed using a sentence transformer model (all-MiniLM-L6-v2). Scores closest to 1 indicate near-identical meaning, whereas values towards zero mean poor similarity.

BERTScore[44] is a semantic evaluation metric that computes token-level similarity between candidate and reference texts using contextual embeddings from BERT. It provides three scores: Precision (how much of the generated text is supported), Recall (how much of the reference is captured), and F1 (harmonic mean). Scores higher than 0.95 represent an excellent semantic match, while lower than 0.80 suggest review may be needed.

See Table 2 for a summary and for target reference values.

### 5.4 Human evaluation

To complement the metrics above, we include a targeted human evaluation of factual accuracy. Since the primary evaluation relies on the eight automated metrics defined above, the human assessment is not intended as a comprehensive review of all 112 outputs, but rather as a focused validation of specific model behaviours flagged by the metrics. It was therefore conducted by a single domain expert, who evaluated factual accuracy against the source documents using an Expert Accuracy Rating (EAR), assessing whether the information generated by LLMs was verifiable in the input material. Cases were selected through strategic sampling, prioritising instances where the metrics signalled the greatest inter-model divergence in placeholder behaviour, the largest discrepancy between strategies with and without content rules, and extreme verbosity or redundancy.

## 6. Results

In the following figures, we present the results as heatmaps. In Figure 7 there are lexical metrics, structural in Figure 8, and semantic in Figure 9.

### 6.1 The gold standard similarity paradox

A noteworthy result in Figure 9 is that S7 (baseline) achieves the highest Gold Standard Similarity for both models (0.749 for qwen3-32b, 0.753 for coder-30b), while S1 (full prompt) scores lower (0.727 and 0.733 respectively). S5 (free-form) follows a similar pattern. This apparent paradox has a methodological explanation: Gold Standard Similarity computes

**Table 2.** **Metrics in use and target values.**

| Metric | Formula | Target/ Range |
|---|---|---|
| Flesch-Kincaid Grade | FK Grade = 0.39 × (total words / total sentences) + 11.8 × (total syllables/ total words) − 15.59 | 10–14 |
| Flesch Reading Ease | FRE = 206.835–1.015 × (total words / total sentences) − 84.6 × (total syllables/total words) | 30–50 |
| Type-Token Ratio | TTR = (Number of unique words) / (Total number of words) | 0.5–0.8 |
| Redundancy Score | Redundancy = 1 − (Number of unique bigrams/Total number of bigrams) | 0–0.2 |
| Template Compliance | TCS = (Number of present components) / (Total required components) | ≥ 0.8 |
| Placeholder Ratio | PR = (Number of placeholders found) / (Expected fields) | 0 |
| Gold Standard Similarity | Similarity = cos(θ) = (E_gen · E_ref) / (\|\|E_gen\|\| × \|\|E_ref\|\|) | ≥ 0.8 |
| BERTScore F1 | F1 = 2 × (P × R)/(P + R) | ≥ 0.85 |

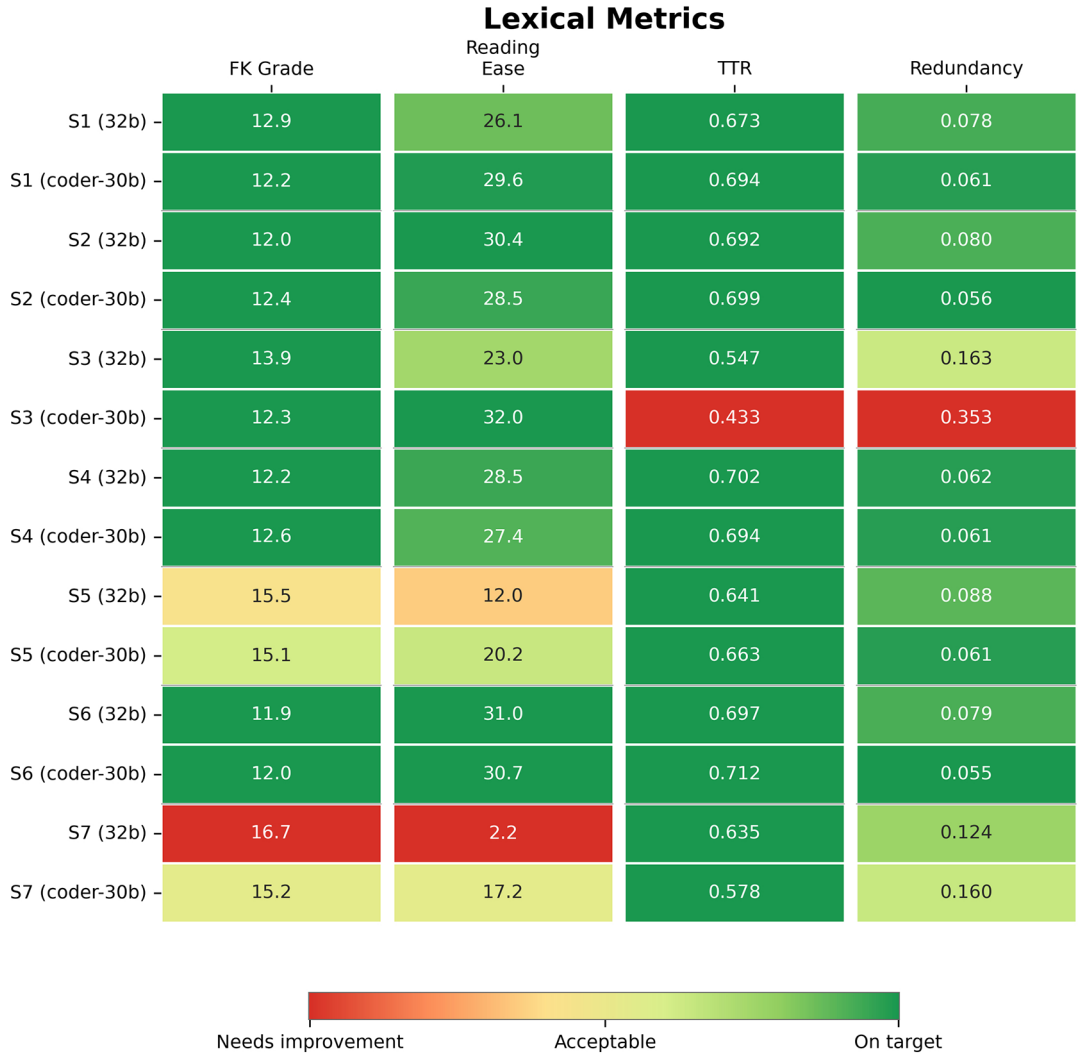



**Figure 7.** **Results: lexical metrics.**




**Figure 8. Results: structural metrics.**

cosine similarity between sentence embeddings, which captures overall semantic proximity but is sensitive to surface-level phrasing. The gold standard descriptions were written by human experts in natural prose, and the unconstrained strategies (S5, S7) produce text that more closely resembles such writing style at the embedding level. Template-guided strategies, by contrast, impose a rigid structure ("This dataset provides… It depicts…") that diverges from the gold standard's phrasing even when conveying equivalent information.

BERTScore F1 produces a contrasting ranking: S1 and S4 in Figure 9 score highest (~0.836), while S7 scores lowest (0.816 for qwen3-32b, 0.824 for coder-30b). This can be explained considering that BERTScore operates at token level, matching individual words and phrases across texts, making it more robust to structural rearrangement. It captures that the template-guided outputs contain the same informational content as the gold standard, even if it is organised differently. The divergence between the two semantic metrics underscores the importance of employing complementary measures: Gold Standard Similarity reflects stylistic and structural alignment, while BERTScore reflects informational overlap regardless of form.

### 6.2 Template compliance – clear ablation signal

The TCS results provide a validation of the ablation design. In Figure 8, strategies that include the use of a template (S1, S2, S4, S6) consistently achieve TCS scores between 0.95 and 0.99, indicating that nearly all 8 expected components are present. Strategies with no template provided (S5,S7), or with template provided but lacking rules/examples (S3), score 0.80 or lower across both models. The component-level breakdown reveals the source of this gap: the template-specific

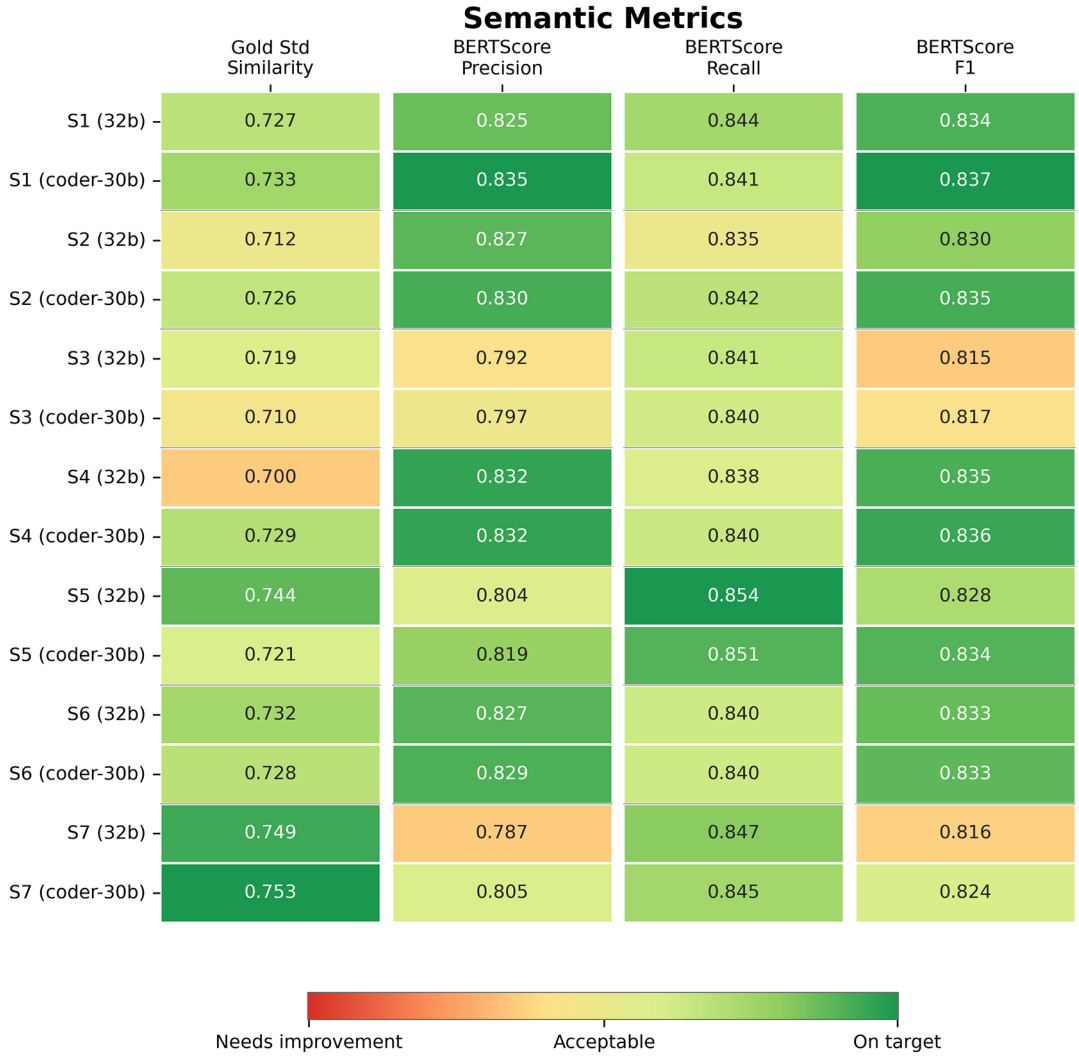



**Figure 9. Results: semantic metrics.**

opening phrases ("This dataset provides" and "It depicts") are absent in all eight datasets for every non-template strategy, confirming that these structural elements are never produced spontaneously by both models. This binary distinction - near-complete compliance with the template versus systematic absence of its signature phrases - constitutes a clean ablation signal, demonstrating that the template is the single most influential layer in determining structural output quality.

### 6.3 Model difference - placeholder behaviour

The most notable inter-model difference concerns Placeholder Ratio. In Figure 8, across template-guided strategies (S1, S2, S4, S6), qwen3-32b averages 0.05–0.11, while qwen3-coder-30b has near-zero PR (0.00–0.02) on the same strategies. On the surface, the coder model appears to provide more complete results. However, a lower Placeholder Ratio is not necessarily indicative of higher quality - it may instead reflect the model's tendency to avoid acknowledging information gaps.

To investigate this, we selected the three cases with the largest inter-model placeholder discrepancy and compared outputs side by side against the source documents: Dataset 7/ S2 (5 placeholders in qwen3-32b vs 0 in coder), Dataset 3/ S1 (4 vs 0), and Dataset 4/S6 (3 vs 0).

The manual inspection revealed that the coder model employs two distinct strategies to avoid placeholders. The first is **omission**: the coder produces shorter output than qwen3-32b by silently dropping sentences where information is

unavailable, rather than flagging them with "[Not specified]." This pattern was observed in Dataset 3/S1, Dataset 4/S1, Dataset 5/S2, Dataset 5/S6, and Dataset 8/S1. The second strategy is **fabrication**: the coder produces output of equal or greater length, replacing placeholders with plausible-sounding content that cannot be found in the source documents. For instance, in Dataset 4/S6, the coder wrote "statistical multivariate similarity" and "sourced from the Joint Research Centre" in fields where qwen3-32b used "[Not specified]."

Both behaviours are problematic from a metadata quality perspective. Omission is arguably worse than a placeholder, as it leaves the user unaware that information is missing. Fabrication introduces factual inaccuracies into a metadata record that the user has no reason to question. By contrast, qwen3-32b's use of explicit placeholders, while producing a less polished output, preserves transparency about the limits of what can be extracted from the source documents. This finding has practical implications for production deployment: a model that seems to generate complete metadata may in fact require more-not less-human oversight.

Recent research on instruction-tuned LLMs provides an explanation for these behaviours. Zhang et al.[45] demonstrate that standard instruction tuning forces models to complete responses regardless of knowledge availability, causing them to fabricate rather than indicate uncertainty. Zhu et al.[46] show that reliable refusal behaviours requires specialized Refusal-Aware Instruction Tuning (RAIT), which standard instruct models lack. Jha et al.[47] confirm that Qwen models trained without explicit abstention rewards prefer to generate or delete text rather than output uncertainty markers. These findings suggest that the coder model's placeholder avoidance is a predictable consequence of instruction tuning optimized for helpfulness over epistemic honesty-a design choice that, while understandable, proves counterproductive for metadata generation tasks where transparency about missing information is essential.

### 6.4 S3 (No examples, no template rules) - coder model shows high redundancy

In Figure 7, the coder model on S3 has a redundancy value equal to 0.35 - nearly 6x higher than the template strategies (~0.06). Without examples and template rules to guide it, the coder model becomes repetitive. The general model handles this better (0.16). It suggests that the coder model can be more dependent on structured prompting guidance.

In terms of words count, S3/coder is a clear outlier, scoring 3.4 times more than the global word count average (209.7 words), driven by two extreme cases (Datasets 7 and 8) where the model produced over 2000 words with redundancy exceeding 0.82. This suggests that the coder model is more dependent on explicit structural constraints - when extraction rules and few-shot examples are removed, it defaults to repetitive generation rather than concise extraction. The general-purpose model, by contrast, maintains comparable output length across constrained and unconstrained strategies.

### 6.5 Readability – S5 and S7 are less readable

Strategies without a template produce noticeably less readable text. In Figure 7, S5 (free-form) and S7 (baseline) exhibit Flesch-Kincaid Grade Levels of 15–17 and Flesch Reading Ease scores between 2 and 20, placing them well beyond the range considered appropriate for technical documentation (FK 10–14, FRE 30–50). By contrast, all template-guided strategies (S1, S2, S4, S6) remain within this target range, with FK Grade around 12 and FRE between 26 and 31. This suggests that the template serves a dual function: in addition to enforcing structural compliance, it constrains sentence complexity and length, resulting in more accessible text. Without this guidance, both models default to longer, more syntactically complex sentences that could be difficult to parse even to expert readers.

### 6.6 Layers impact

The ablation analysis reveals a clear hierarchy of layer impact. The template emerges as the most influential single layer: beyond enforcing structural compliance, it constrains readability within acceptable bounds for technical documentation (FK Grade ~ 12, FRE 26–31) and reduces redundancy by defining a finite set of fields after which the model stops generating.

Few-shot examples, combined with detailed extraction rules, constitute the second most impactful layer, as evidenced by the transition from S3 (template without rules or examples) to S6 (template with rules and positive examples): redundancy drops from 0.258 (cross-model average $\approx(0.35 + 0.16)/2$) to 0.067, TTR improves from 0.490 to 0.704, and BERTScore F1 increases from 0.816 to 0.833. While this comparison does not isolate the individual contribution of examples from that of rules, the magnitude of the improvement - particularly in redundancy - suggests that examples play a substantial role in calibrating the expected output length and style. Without this guidance, particularly the coder model defaults to repetitive, verbose generation.

Content rules represent a more subtle but functionally important layer: they do not measurably improve output quality, but they increase the model's tendency to acknowledge information gaps through explicit placeholders rather than generating potentially fabricated content.

The expert role definition contributes the least measurable effect of any individual layer, with no metric showing a difference greater than 0.01 between S1 (role on) and S2 (role off).

Finally, model choice interacts significantly with the other layers: both models perform comparably under full prompting guidance, but diverge markedly in the absence of structural constraints - the coder model producing extreme verbosity under S3 and employing omission or fabrication strategies to avoid placeholders where the general-purpose model explicitly flags missing information.

## 7. Discussion and conclusions

This study investigated the feasibility of using Large Language Models (LLMs) to assist in generating structured metadata descriptions for geospatial datasets compliant with the GeoDCAT-AP schema. Through a systematic ablation of six prompting layers across seven strategies and two LLMs, we identified a clear hierarchy of component impacts on output quality.

Based on the findings of this study, we provide recommendations for automated metadata description generation in what follows.

Strategy S1 (full expert configuration) is recommended for production use, as it achieves the highest BERTScore F1 (0.836), near-maximum Template Compliance (0.983), and -crucially- incorporates content rules that encourage the model to acknowledge information gaps rather than fabricate content. While its Placeholder Ratio is slightly higher than other strategies, this reflects transparency rather than deficiency: explicit "[Not specified]" markers are to be preferred to silently omitted or fabricated information in a metadata production context, where downstream users and systems may rely on the description as authoritative.

Strategy S6 (positive examples only, without negative examples) represents a viable lighter alternative when prompt token budget is a concern. Dropping negative examples reduces prompt length while incurring negligible quality loss: TCS decreases by only 0.012, BERTScore F1 by 0.003, and Gold Standard Similarity remains identical. This makes S6 suitable for deployment scenarios involving models with smaller context windows or where API costs are proportional to input token count.

Strategy S4 (no content rules) should be avoided despite producing outputs that appear more complete. Its lower Placeholder Ratio (0.058 vs 0.100 for S1) masks the absence of a hallucination safeguard: without explicit instructions to extract only information present in the source documents, the model is more likely to fill template fields with plausible but unverifiable content. Manual inspection confirmed instances where S4 generated content for fields that S1 correctly flagged as unavailable in the source material.

Regarding model selection, both qwen3-32b and qwen3-coder-30b-a3b-instruct produce semantically comparable outputs under full prompting guidance. However, qwen3-32b exhibits more conservative behaviour when dealing with missing information, leaving placeholders where the coder model either omits or fabricates content. For metadata generation tasks, where factual accuracy is paramount, the general-purpose model is therefore preferable. The coder model may still be suitable in workflows where a human reviewer systematically verifies every generated description, but its tendency toward silent omission and subtle fabrication makes it a riskier choice for semi-automated pipelines with limited human oversight.

More broadly, this study demonstrates that prompt engineering for metadata generation is not merely a matter of maximising output quality - it requires balancing completeness, structural compliance, and factual honesty. The ablation design reveals that each prompting layer serves a distinct function, and that removing even seemingly minor layers (such as content rules) can compromise the reliability of the output in ways that are not captured by standard automated metrics alone.

### 7.1 The structure-honesty trade-off

The recommendation to avoid S4, despite its apparent completeness, reflects a deeper tension in LLM-assisted generation: the techniques that improve output quality may also create conditions for hallucinations. Template-guided

prompting enforces structural compliance and readability, but it also creates expectations that models may feel compelled to satisfy, even if source documents lack specific information.

Automated evaluation metrics that reward completeness may inadvertently favour outputs that fabricate rather than abstain. It is a risk that human evaluation confirmed (see Section 5.4): in fact, manual inspection revealed fabricated content in fields where source documents lacked information.

Research on Refusal-Aware Instruction Tuning[45–47] provides theoretical grounding for this observation. Standard instruction tuning creates models that treat every prompt as answerable -a design assumption that conflicts with metadata generation- where acknowledging "*information not available in source*" is often the correct response. Our content rules layer functions as a prompt-level workaround for this architectural limitation, encouraging the model to leave placeholders rather than fabricate. However, this safeguard is inherently fragile and requires careful (and thus costly) monitoring in production deployments.

### 7.2 Implications for semi-automated workflows

Our findings challenge the assumption that better-performing models require less human oversight. The lower Placeholder Ratio of the coder model initially suggested superior performance, but this metric conflates genuine completeness with concealed gaps. In practice, a model that explicitly marks missing information makes the reviewer's task clearer: the human editor knows exactly which fields require attention. A model that silently omits or fabricates content provides no such guidance, potentially requiring verification of every field against source documents. This distinction is particularly relevant for semi-automated pipelines where a comprehensive manual review of every output is likely not doable.

### 7.3 Addressing metadata challenges

The proposed approach directly addresses two challenges in metadata creation identified in Section 3: the difficulty of generating appropriate free-text content (Section 3.2.1) and the blank-sheet syndrome (Section 3.2.2). By providing draft content that human editors can refine rather than creating from scratch, the approach is suited for scenarios where data volumes are considerable and metadata updates are frequent, such as Common European Data Spaces, digital libraries, and Spatial Data Infrastructures. The template encodes decisions about what information should appear, in what order, and with what level of detail-decisions that the LLM then applies consistently across outputs, promoting harmonisation across metadata records produced by different operators or at different times.

While we focused on geospatial data and the GeoDCAT-AP schema, the methodology may be applicable to other domains. The classification of attributes into Types A (descriptive), B (technical), and C (institutional) is not specific to GeoDCAT-AP; similar structures appear across metadata standards.

### 7.4 Limitations

This study has limitations that should be acknowledged. The evaluation relied primarily on automated metrics; while targeted manual inspection confirmed fabrication and omission behaviours, a comprehensive factual accuracy assessment across all 112 outputs was not carried out. Both evaluated models belong to the Qwen family, and the observed behaviours -particularly placeholder avoidance in the coder model- should not be generalised to other architectures and models.

Additionally, all eight datasets originate from a single catalogue (the JRC Data Catalogue), which, despite internal variability in documentation practices, may not fully capture the diversity of metadata conventions found across different organisations and domains.

## Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used:

- GPT@JRC (internal JRC deployment of LLMs) in order to: 1) generate the metadata, as presented in the article; 2) improve readability of the text
- Claude AI (Anthropic) to help generate python scripts for the evaluation suite

After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## Software availability

Analysis code available from: https://code.europa.eu/jrc-dsi-lab/metadata_evaluation_suite

Archived analysis code as at time of publication: https://doi.org/10.5281/zenodo.20623178 (Di Leo, 2026b).

License: European Union Public License 1.2 (EUPL 1.2).

## Ethics and consent

Ethical approval and consent were not required.

## Data and software availability

### Extended data

Extended data are published on Zenodo at https://doi.org/10.5281/zenodo.20623178 (Di Leo, 2026a).[48]

The same Zenodo deposit contains the code snapshot together with the input and output data for full reproducibility (Di Leo, 2026b)[49]. The datasets are licensed under CC-BY 4.0.

Hereinafter is a description of the individual files.

- *ImportantProperties.xlsx*: mapping of the attributes of the GeoDCAT-AP schema into Types A, B and C; values of the attributes for the dataset used in the study.
- *Results_paper.xlsx*: the 112 generated descriptions
- *metrics_results.xlsx*: the generated metrics
- *Metadata_metrics.py*: the python script to generate the metrics. Takes in input *Results_paper.xlsx* and generates *metrics_results.xlsx* as output
- *Heatmap_metrics.py*: the python script used to generate heatmaps of the metrics. Takes *metrics_results.xlsx* as input and generates the heatmaps in png and svg
- Prompts folder: the folder containing the prompts used to generate the descriptions (in markdown)
- Datasets folder: folder containing all information sourced for each dataset.

## Acknowledgements

The authors would like to thank Alexander Kotsev and Anders Friis-Christensen (JRC) for the continuous support and valuable input provided during the work.

## References

1. Bode J, Kühl N, Kreuzberger D, *et al.*: **Toward avoiding the data mess: industry insights from data mesh implementations.** *IEEE Access.* 2024; **12**: 95402–95416.
Publisher Full Text

2. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The fair guiding principles for scientific data management and stewardship.** *Scientific data.* 2016; **3**(1): 1–9.
PubMed Abstract | Publisher Full Text | Free Full Text

3. **Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an infrastructure for spatial information in the European community (INSPIRE).**2007.

4. *Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast) OJ L 172, 26.6.2019.* ELI;56–83.
Reference Source

5. *Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) OJ L 152, 3.6.2022.* ELI;1–44.
Reference Source

6. Driessen SW, Monsieur G, Van den Heuvel WJ: **Data market design: A systematic literature review.** *IEEE access.* 2022; **10**: 33123–33153.
Publisher Full Text

7. Bendechache M, Attard J, Ebiele M, *et al.*: **A systematic survey of data value: models, metrics, applications and research challenges.** *IEEE Access.* 2023; **11**: 104966–104983.
Publisher Full Text

8. Stow CA, Webster KE, Wagner T, *et al.*: **Small values in big data: The continuing need for appropriate metadata.** *Ecological*

*Informatics.* 2018; **45**: 26–30.
**Publisher Full Text**

9. **Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A European strategy for data.**2020.
**Reference Source**

10. Ulrich P, Duch Brown N, Kotsev A, *et al.*: *Quantifying the Benefits of Location Interoperability in the European Union, EUR 31004 EN.* Luxembourg: Publications Office of the European Union; 2022. JRC127330. 978-92-76-48846-0.
**Publisher Full Text**

11. Bagchi M: **A Generative AI-driven Metadata Modelling Approach. arXiv preprint arXiv:2501.04008.** 2025.
**Publisher Full Text**

12. Yang W, Fu R, Amin MB, *et al.*: **The impact of modern AI in metadata management.** *Human-Centric Intelligent Systems.* 2025; **5**: 323–350.
**Publisher Full Text**

13. European Parliament and European Council: **Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).** *Official Journal of the European Union, L.* 2007; **108**: 1–14.
**Reference Source**

14. Chou C, Chu T: **An analysis of BERT (NLP) for assisted subject indexing for project Gutenberg.** *Cataloging & Classification Quarterly.* 2022; **60**(8): 807–835.
**Publisher Full Text**

15. El Ganadi A, Vigliermo RA, Sala L, *et al.*: **Bridging Islamic knowledge and AI: Inquiring ChatGPT on possible categorizations for an Islamic digital library (full paper).** *CEUR Workshop Proceedings.* CEUR-WS; 2023; Vol. **3536**: pp. 21–33.
**Reference Source**

16. Martorana M, Kuhn T, Stork L, *et al.*: **Zero-shot topic classification of column headers: Leveraging llms for metadata enrichment.** *arXiv preprint arXiv:2403.00884.* 2024.
**Publisher Full Text**

17. Zhang S, Wu M, Zhang X: **Utilising a large language model to annotate subject metadata: A case study in an Australian national research data catalogue. arXiv preprint arXiv: 2310.11318.** 2023.
**Publisher Full Text**

18. Zhang H, Liu Y, Santos A, *et al.*: **AutoDDG: Automated Dataset Description Generation using Large Language Models.** 2025.
**Publisher Full Text** | **Reference Source**

19. Nair A, Goh ZR, Liu T, *et al.*: **Web Archives Metadata Generation with GPT-4o: Challenges and Insights.** *Information Technology and Libraries.* 2025; **44**(2).
**Publisher Full Text**

20. Lu Z, van der Plas TL, Rashidi P, *et al.*: **Flexible metadata harvesting for ecology using large language models.** *International Conference on Theory and Practice of Digital Libraries.* Cham: Springer Nature Switzerland; 2025, September; pp. 338–352.
**Publisher Full Text**

21. Zavalin V, Zavalina OL: **Are we there yet? Evaluation of AI-generated metadata for online information resources.** *Information Research an international electronic journal.* 2025; **30**(iConf): 732–740.
**Publisher Full Text**

22. Ulrich H, Kock-Schoppenhauer A, Deppenwiese N, *et al.*: **Understanding the Nature of Metadata: Systematic Review.** *J. Med. Internet Res.* 2022; **24**(1): e25440.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

23. Batista D, Gonzalez-Beltran A, Sansone SA, *et al.*: **Machine actionable metadata models.** *Scientific Data.* 2022; **9**(1): 592.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

24. Martínez-García M, Hernández-Lemus E: **Data integration challenges for machine learning in precision medicine.** *Frontiers in medicine.* 2022; **8**: 784455.
**Publisher Full Text**

25. Kim J, Yakel E, Faniel IM: **Exposing standardization and consistency issues in repository metadata requirements for data deposition.** *College & Research Libraries.* 2019; **80**(6): 843–875.
**Publisher Full Text**

26. **International Organization for Standardization, ISO 19115:2014 – Geographic information - Metadata.**2014. (accessed 30 September 2025).
**Reference Source**

27. **SEMIC, GeoDCAT-AP – A DCAT Application Profile for spatial datasets and services, version 3.1.0.**2026. (accessed 20 February 2026).
**Reference Source**

28. Ureña-Cámara MA, Nogueras-Iso J, Lacasta J, *et al.*: **A method for checking the quality of geographic metadata based on ISO 19157.** *International Journal of Geographical Information Science.* 2019; **33**(1): 1–27.
**Publisher Full Text**

29. Patil R, Boit S, Gudivada V, *et al.*: **A survey of text representation and embedding techniques in NLP.** *IEEE Access.* 2023; **11**: 36120–36146.
**Publisher Full Text**

30. World Wide Web Consortium: *Data Catalog Vocabulary (DCAT) – Version 3.* W3C Recommendation; 2024. (accessed 20 October 2025).
**Reference Source**

31. DCMI: **Metadata Terms, Term name: description, Tech. rep.**
**Reference Source**

32. Joint Research Centre, Joint Research Centre Data Catalogue, European Commission: 2025. (accessed 30 September 2025).
**Reference Source**

33. OpenAI: *Hello GPT-4o.* OpenAI; 2025. (accessed 30 September 2025).
**Reference Source**

34. Sai AB, Mohankumar AK, Khapra MM: **A survey of evaluation metrics used for nlg systems.** *ACM Computing Surveys (CSUR).* 2022; **55**(2): 1–39.
**Publisher Full Text**

35. Kincaid JP, Fishburne RP, Rogers RL, *et al.*: *Derivation of New Readability Formulas for Navy Enlisted Personnel. Research Branch Report 8-75.* Naval Technical Training Command; 1975.

36. Flesch R: **A new readability yardstick.** *Journal of Applied Psychology.* 1948; **32**: 221–233.
**PubMed Abstract** | **Publisher Full Text**

37. Richards B: **Type/Token Ratios: What do they really tell us?.** *Journal of Child Language.* 1987; **14**(2): 201–209.
**PubMed Abstract** | **Publisher Full Text**

38. Fu Z, Lam W, So AM-C, *et al.*: **A Theoretical Analysis of the Repetition Problem in Text Generation.** *Proceedings of AAAI Conference on Artificial Intelligence.* 2021; **35**: 12848–12856.
**Publisher Full Text**

39. Bruce TR, Hillmann DI: *The continuum of metadata quality: defining, expressing, exploiting.* ALA editions; 2004.

40. Ochoa X, Duval E: **Automatic evaluation of metadata quality in digital repositories.** *International journal on digital libraries.* 2009; **10**(2): 67–91.
**Publisher Full Text**

41. Qin Y, Song K, Hu Y, *et al.*: **Infobench: Evaluating instruction following ability in large language models.** *arXiv preprint arXiv: 2401.03601.* 2024.

42. Neumaier S, Umbrich J, Polleres A: **Automated quality assessment of metadata across open data portals.** *Journal of Data and Information Quality (JDIQ).* 2016; **8**(1): 1–29.
**Publisher Full Text**

43. Reimers N, Gurevych I: **Sentence-bert: Sentence embeddings using siamese bert-networks.** *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP).* 2019; pp. 3982–3992.

44. Zhang T, Kishore V, Wu F, *et al.*: **BERTScore: Evaluating Text Generation with BERT.** *International Conference on Learning Representations (ICLR).* 2020. arXiv:1904.09675.

45. Zhang H, Diao S, Lin Y, *et al.*: **R-Tuning: Instructing Large Language Models to Say 'I Don't Know'.** *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).* 2024; pages 7113–7139.
**Publisher Full Text**

46. Zhu R, Jiang Z, Wu J, *et al.*: **GRAIT: Gradient-Driven Refusal-Aware Instruction Tuning for Effective Hallucination Mitigation.** *Findings of the Association for Computational Linguistics: NAACL 2025.* 2025; pages 4006–4021.
**Publisher Full Text**

47. Jha A, Mahajan A, Aravindan AV, *et al.*: **Rewarding Intellectual Humility: Learning When Not to Answer in Large Language Models.** *arXiv preprint arXiv:2601.20126.* 2026.

48. Di Leo M: Extended data linked to paper: Towards AI-Assisted Metadata Generation For Improved Description of Geospatial Data. [Data set]. *Zenodo.* 2026a.
**Publisher Full Text**

49. Di Leo M: **Extended data linked to paper: Towards AI-Assisted Metadata Generation For Improved Description of Geospatial Data [Software].** *Zenodo.* 2026b.
**Publisher Full Text**

# Open Peer Review

## Current Peer Review Status: ?

**Version 1**

Reviewer Report 11 August 2026

https://doi.org/10.21956/openreseurope.26284.r78878

© **2026 Sherman Z.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

**Zachary Sherman**
Department of Geography, Virginia Tech, Blacksburg, Virginia, USA

**Summary of the article**
This paper investigates whether Large Language Models (LLMs) can help human data producers create GeoDCAT-AP-compliant dataset descriptions, following a human-in-the-loop approach and motivated by the cost of manual metadata curation and the EU data-space policy context. The authors contribute a three-phase metadata lifecycle (generation, exposure, exploitation), a taxonomy of GeoDCAT-AP attributes (descriptive, technical, institutional), and a six-layer configurable prompting framework. They run a layered ablation across seven prompt configurations (S1–S7), two Qwen models (general-purpose and code-specialised), and eight geospatial datasets, targeting the free-text description attribute. The 112 outputs are assessed with eight automatic metrics (lexical, structural, semantic) plus a targeted expert factual-accuracy check.

The main findings are a clear hierarchy of prompt-layer impact, the input template is most influential (structural compliance and readability), few-shot examples plus rules are second (reducing redundancy), content rules act as an anti-fabrication safeguard rather than a quality booster, and expert role definition has the least effect. A valuable qualitative result is the structure–honesty trade-off: the code-specialised model achieves a near-zero placeholder ratio by silently omitting or fabricating content, whereas the general-purpose model transparently flags missing information, so a lower placeholder ratio can indicate worse, not better, metadata.

**General appraisal**
This is a clear, well-motivated, and practically relevant study, and the ablation design is sound. The structure–honesty finding is a genuinely useful, deployment-oriented contribution that goes beyond the expected "prompt engineering helps" result. The recommendations for production use are sensible. The revisions below are intended to sharpen the positioning and firm up the evidence behind a few of the stronger claims; none require re-running the core experiment.

Expanded responses to the review questions

**1. Clear presentation and current literature — Partly.**
The writing and framework are clear. Two adjustments would help: (a) the novelty statement ("not covered in the literature to the best of our knowledge") is a little too strong given several adjacent works, and would be more accurate if reframed as a precise gap, systematic prompt-layer ablation for the GeoDCAT-AP description attribute against human gold-standard references; and (b) a few directly relevant works could be engaged, notably LLMs for FAIR geospatial data (Kavhu et al., *Scientific Data* 2026), geospatial metadata automation in Dataverse (*Environmental Modelling & Software* 2025, doi:10.1016/j.envsoft.2025.106792), LLM extraction of DCAT-compatible metadata (Tebernum et al., doi:10.5220/0013458500003967), AutoDDG for dataset-description generation (doi:10.1145/3786626), and "Faithful or Findable? Evaluating LLM-Generated Metadata for RDF Dataset Search" (arXiv:2607.05970), which reaches a conclusion closely parallel to the authors' own trade-off. The "role definition contributes least" result is also well supported by existing work (e.g., Zheng et al., "When 'A Helpful Assistant' Is Not Really Helpful," EMNLP Findings 2024) and citing it would strengthen the point.

**2. Study design and merit — Yes.**
The six-layer decomposition and seven controlled configurations are a principled way to isolate component contributions, and separating template design (GPT-4o) from generation (Qwen) sensibly avoids circularity. No changes required.

**3. Methodological detail for replication — Partly.**
To enable replication, please add: model inference settings (temperature/decoding parameters, context configuration); the verbatim prompt text for each layer and strategy; and a fuller description of the human-evaluation protocol (the EAR scale/rubric and the exact case-selection procedure). The embedding and BERTScore model versions used would also help.

**4. Statistical analysis — Partly.**
The large effects (template presence; few-shot/rules on redundancy) are convincingly demonstrated. Some finer-grained rankings, however, rest on very small differences reported without dispersion (e.g., S6 vs S1 of TCS −0.012 and BERTScore F1 −0.003; the role-layer "no difference greater than 0.01"), which with eight datasets may be within noise. Please add per-dataset variance and a simple significance or effect-size check, and soften any ranking claim that does not survive it. The divergence between the two semantic metrics is better described as an expected consequence of embedding-level versus token-level similarity rather than a "paradox," with a note that none of the automatic metrics measures factual accuracy (which is why the expert check exists).

**5. Source data availability — Partly.**
Please ensure the Extended Data release includes the eight datasets and supporting corpora (or pointers where licensing prevents redistribution), the gold-standard descriptions, all 112 generated outputs, the complete per-dataset metric values behind the heatmaps, and the human-evaluation ratings, under a persistent identifier.

**6. Conclusions supported by results — Partly.**
The core conclusions are well supported and should be retained: the template is the dominant layer, few-shot examples plus rules reduce redundancy, and content rules safeguard factual honesty (consistent with the refusal-aware/abstention literature the authors cite). Three claims would benefit from light rescoping:

(a) the fine-grained layer ranking based on sub-0.01 differences (pending the statistics above); (b) the general-purpose-versus-code-specialised conclusion, which is drawn from two models of a single family (Qwen) and is best scoped to the models tested unless a further family is added; and (c) the factual-reliability claims, which rest on a single expert reviewing a deliberately selected subset — either add a second annotator on those same cases with inter-rater agreement, or present the finding as indicative single-rater inspection. It would also be worth noting that, because the template was induced from JRC descriptions, high template-compliance partly reflects conformity to a JRC stylistic convention.

Minor points

- The title/abstract imply "metadata generation" broadly, while the study evaluates one attribute (description); a brief scoping note early on would align expectations.
- The S6 "lighter alternative" / cost argument would be more convincing with prompt token counts per strategy.
- The manuscript supplied ends mid-sentence in Section 7.1; please confirm the full Discussion is present.
- Tables 1–2 and the results heatmaps were not fully legible; please ensure readable, machine-readable values in the typeset version and supplement.

Points that should be addressed to make the article scientifically sound

1. Reframe the novelty claim precisely and engage the directly-adjacent geospatial-metadata and dataset-description literature noted above.
2. Add dispersion and a simple significance/effect-size check so the smaller layer-ranking differences are not over-interpreted; reframe the metric "paradox" and note that no automatic metric measures factual accuracy.
3. Report inference settings and release the verbatim prompts, the full per-dataset metric matrix, the datasets/corpora, gold-standard references, all 112 outputs, and the human-evaluation ratings.
4. Either add a second annotator (with inter-rater agreement) on the sampled cases or present the factual-reliability claim as indicative single-rater inspection.
5. Scope the general-purpose-versus-code-specialised conclusion to the tested Qwen models, or add a model from another family.

These are constructive and, for the most part, straightforward to address within a revision cycle; the underlying study is a solid and welcome contribution.

Great Work !

**References**

1. Trišović A, Range J, Durbin P, Mika K, et al.: Advancing geospatial data infrastructure in Dataverse via metadata automation, interactive tools and LLM case study. *Environmental Modelling & Software*. 2026; **199**. Publisher Full Text
2. Zhang H, Liu Y, Santos A, Hung W, et al.: AutoDDG: Automated Dataset Description Generation using Large Language Models. *Proceedings of the ACM on Management of Data*. 2026; **4** (1): 1-27 Publisher Full Text

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and does the work have academic merit?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**
Partly

**Are all the source data underlying the results available to ensure full reproducibility?**
Partly

**Are the conclusions drawn adequately supported by the results?**
Partly

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Generative AI and large language models for geospatial analysis; GeoAI; spatial data science / GIS.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**