Lost and Found: Enriching Knowledge Graphs with "NIL" Persons from Historical Documents

Journal Title
XX(X):1–19
©The Author(s) 2025
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Arianna Graciotti[1], Nicolas Lazzari[2], Enrico Daga[3] and Valentina Presutti[1]

Abstract

Vast community-driven knowledge graphs (KGs), like Wikidata, are the primary reference data for Entity Linking (EL) applications. However, they exhibit significant coverage bias towards information that is widely popular on the Web, leading to underrepresentation of long-tail entities, particularly from non-contemporary contexts. Concurrently, ongoing mass digitisation of cultural heritage resources reveals numerous named entities and associated knowledge currently missing from general-purpose KGs. Enriching such KGs with these "NIL" entities presents an opportunity to enhance their completeness and mitigate biases, such as gender disparity in the representation of historical figures. In this article, we investigate an approach based on retrieval-augmented generative AI to capture information about NIL entities and generate structured KGs suitable for integration into Wikidata. The approach is applied to the case of persons unknown to Wikidata but mentioned in a collection of musical periodicals of the 19th century. We empirically select 6 properties used in Wikidata on entities of that type and create a manually annotated NIL-entities KG as a gold standard for evaluation. Through comprehensive experiments, we evaluate 6 State-of-the-Art Large Language Models (LLMs) from different vendors, combined with 6 different State-of-the-Art retrievers. Our results demonstrate significant variations in performance across model-retriever combinations, with a high accuracy for gender identification and family name, promising results for occupation and country of citizenship, and low for date of birth. We report on a detailed error analysis and discuss the potential of our approach for mitigating historical bias in Wikidata.

1 Introduction

Knowledge graphs (KGs) Hogan et al. (2021) have become key resources for structuring and integrating human knowledge. However, general-purpose, community-driven KGs, such as DBpedia or Wikidata, suffer from data sparsity and coverage biases, particularly in relation to long-tail knowledge such as knowledge embedded within historical documents, domain-specific corpora, recently digitized, or offline textual resources. Cultural heritage is especially affected by this problem. A notable gap concerns historical entities, especially people, mentioned in specialized heritage corpora, such as archival records or scholarly publications Ripoll et al. (2025); Laouenan et al. (2022). These entities are often missing or poorly described in large-scale, open KGs. In addition, due to their limited presence on the open web, they are also underrepresented in Large Language Models (LLMs), which are typically trained on more recent and popular sources. Such a poor representation results in lower performance of state-of-the-art Entity Linking (EL) methods on historical documents Graciotti et al. (2024, 2025b); Arora et al. (2024); Ehrmann et al. (2022).

This situation creates a compounded bias: a popularity bias that privileges well-known contemporary figures, and a structural bias that marginalizes less documented groups - such as women in historical records — whose trace in digital knowledge resources is already limited Graciotti et al. (2024, 2025b). The lack of structured representations for these lesser-known entities limits the scope of digital humanities research and weakens the potential of LLMs and KG-based systems to support diverse and inclusive knowledge applications.

Recent studies Graciotti et al. (2025a); Heist and Paulheim (2023); Dong et al. (2023) have proposed methods to identify missing entities (i.e. NIL entities) in Wikidata. These approaches provide evidence supporting the existence of this gap, but do not address how to enrich Wikidata with these entities and their associated properties in a coherent and reusable way. Bridging this gap is relevant in contexts such as digitization of historical archives or the training of LLMs to handle less popular or diachronic knowledge, thus mitigating hallucinations and enhancing representation of marginalized perspectives.

In this paper, our aim to address the challenge of automatically enriching Wikidata with entities currently absent from it. We experiment on entities of type person (Wikidata's `Q5`) — arguably the most populous class in Wikidata and the richest in diversity of property types. Although the method we investigate is not hardcoded to people, making it applicable to any other entity type, we concentrate on this entity type for its expressive property structure and relevance to our challenge. Our approach consists in investigating Retrieval Augmented Generation

[0][1]LILEC, University of Bologna, Italy
[2]Computer Science Department, University of Pisa, Italy
[3]Knowledge Media Institute, The Open University, United Kingdom
[0]
Corresponding author:
Arianna Graciotti, arianna.graciotti@unibo.it

(RAG) for knowledge extraction. We start with an historical corpus and a set of entities of a given type derived from it, which are absent in Wikidata. In order to identify the information need related to that type of entity, we select the set of most popular properties used in Wikidata on entities of that type, with a principled approach. Next, for each entity and property pair, we address the information extraction task as a question answering problem in a RAG setting. Through a thorough error analysis, we investigate the interplay between the nature of information need and the parameters of the RAG setting (zero-shot/in-context, retrievers, and language models), showing how there is no single combination that outperform all the others. Therefore, we investigate the impact of orchestrating multiple combinations of the RAG settings. Crucially, our findings challenge the assumption that the performance of LLMs in knowledge extraction require maximum-scale models, showing how mid-range models coupled with retrieval augmentation techniques often outperform larger ones.

We contribute: (1) a manually curated, Wikidata-compliant KG of NIL entities (persons) drawn from historical music periodicals, annotated with six empirically selected core properties; and (2) a generalizable pipeline for KG generation from specialized corpora.

Our experiments, performed on a corpus of periodicals in the music domain of the 19th century, demonstrate the value of the RAG setting for extracting knowledge about NIL entities and show how LLMs can be orchestrated for knowledge enrichment tasks. Furthermore, it contributes toward building more inclusive and historically grounded knowledge infrastructures. The code and data are openly available online[1] under the license CC-BY-SA-4.0.

The rest of the article is structured as follows. Section 2 discusses relevant related research and their relation with our study. In Section 3 the proposed approach to generate KGs for missing entities from specialized corpora is presented, while Section 4 describes the application of our method for enriching Wikidata with historical figures of the music domain, providing details on the experimental setting, the systematic evaluation of six LLMs' performance on this task, and two baseline approaches based on consensus of six LLMs. Results are presented in Section 5 and discussed in Section 6. We conclude by identifying future developments in Section 6.

2 Related Work

This section reviews related work in six areas. Section 2.1 examines KE from historical documents. Section 2.2 covers NIL entity recognition and linking methods. Section 2.3 discusses retrieval-augmented generation for long-tail knowledge. Section 2.4 surveys KG construction from text. Section 2.5 analyzes bias patterns in Wikidata.

2.1 Knowledge Extraction from Historical Documents

The large-scale digitization of historical archives has led to a growing interest in KE from historical documents in recent years, revealing methodological and technological challenges.

Ehrmann et al. (2023) provides a comprehensive overview of the resources for Named Entity Recognition (NER) on historical documents. HIPE-2020 Ehrmann et al. (2020) and HIPE-2022 Ehrmann et al. (2022) initiatives focused on the EL task: HIPE-2022 covers six multilingual (English, Finnish, French, German and Swedish) datasets drawn from 18th to 20th century newspapers and classical commentaries, such as NewsEye Hamdi et al. (2021), SoNAR Menzel et al. (2021), *Le Temps* Ehrmann et al. (2016), the TopRes19th toponym annotations, and the Ajax Multicommentary corpus[2]. Outside of HIPE, Santini et al. (2022) annotates Giorgio Vasari's *Lives of the Artists* with NER, NEC, and links to Wikidata. Santini et al. (2024) annotates Giacomo Leopardi's Zibaldone (1898) for the NER task. Blouin et al. (2024) builds an NER and EL dataset from historical Chinese newspapers (1872-1949) and bilingual biographies from the early twentieth century. Arora et al. (2024) tackles the disambiguation of historical individuals absent from modern KGs by constructing a large-scale dataset from Wikipedia contexts and disambiguation pages, training contrastive bi-encoder models that achieve state-of-the-art linking performance in English-language newswire data, also for NIL entities. KE-MHISTO Graciotti et al. (2025b) constitutes a multilingual (English and Italian) and multitask (NER, EL and QA) benchmark based on historical documents. It includes MHERCL Graciotti et al. (2025a), a manually annotated benchmark of nineteenth-century English-language music periodicals that features a high proportion of underrepresented or missing KG entities. The same work shows that specialised EL systems and LLMs perform poorly on MHERCL, and proposes a KG-enhanced EL approach improved with heuristics to handle NIL predictions, achieving state-of-the-art performances. Building on the KG-enhanced EL method formulated in Graciotti et al. (2025a), Lazzari et al. (2024) addresses the low performance in the EL task due to popularity bias in historical documents by filtering implausible candidates through Answer Set Programming. Using explicit Wikidata knowledge, their method improves retrieval accuracy for less popular entities, demonstrating that structured knowledge can compensate for training data limitations in historical contexts.

Research on QA over historical corpora remains comparatively underexplored. ArchivalQA Wang et al. (2022) presents a large-scale dataset tailored to temporal news archives, categorizing questions by difficulty and temporal expressions to assess the model's ability to retrieve information from decades-old news. Similarly, ChroniclingAmericaQA Piryani et al. (2024) compiles QA pairs from 120 years of U.S. newspapers, evaluating systems on noisy OCR text, human-corrected transcriptions, and raw scanned images to mirror real-world digitization challenges. Unlike typical web-based QA benchmarks, both focus on diachronic language change and temporal reasoning, revealing the limitations of current LLMs on historical data. In the Italian-language context, QUANDHO Menini et al.

[1] https://github.com/arianna-graciotti/KG_Construction_Historical_NIL_Entities/tree/master
[2] https://mromanello.github.io/ajax-multi-commentary/

(2016) offers a QA dataset covering the early 20th century Italian history, complete with manually classified question types, answer pairs and lexical answer-type annotations, which supports domain adaptation for QA systems.

However, none of these historical KE efforts addresses the problem of enriching mainstream KGs such as Wikidata with missing knowledge from historical sources. Our work bridges this gap by leveraging historical texts not only as a source of out-of-KG entities, but also to generate triples suitable for Wikidata or other KGs.

2.2 NIL entities recognition and linking

Most state-of-the-art EL systems assume that every mention can be linked to an existing entry in a knowledge base (KB) of reference, such as Wikidata, effectively ignoring the possibility of out-of-KB entities.

Mainstream benchmarks, such as (Hoffart et al. 2011, AIDA CoNLL-YAGO), (Guo and Barbosa 2018, WNED-WIKI) and (Gabrilovich et al. 2013, WNED-CWEB) focus primarily on in-KB entities providing only limited evaluation possibilities for NIL detection. In contrast, historical EL benchmarks such as HIPE-2022 Ehrmann et al. (2022) and KE-MHISTO Graciotti et al. (2025b) include a substantial proportion of NIL entities.

Early heuristic methods approach the NIL prediction problem using a thresholding function. When the likelihood that a mention should be linked to an entity is less than a threshold, NIL is predicted for that entity. More recent methods train supervised classifiers, often using neural features, to distinguish linkable mentions from NIL cases Sevgili et al. (2022). Building on these ideas, BLINKout Dong et al. (2023) augments a BERT-based linker with a dynamic NIL representation and a learned classifier that predicts more accurately when no valid KG entry exists. However, methods such as ReFinED Ayoola et al. (2022) improve the performance of long-tail and NIL entities by considering KG knowledge in the linking process and treating NIL as a valid linking candidate during their training phase.

Despite the different proposals, none of those methods uses NIL mentions for KG enrichment. Our work builds on these NIL recognition techniques not only by detecting out-of-KG mentions but by constructing triples that can be directly integrated into Wikidata or other knowledge graphs.

2.3 Retrieval-Augmented Generation for Long-Tail Knowledge

Retrieval-augmented generation (RAG) combines LLMs with an external retrieval component Lewis et al. (2020); Gao et al. (2024); Fan et al. (2024) grounding the output to external knowledge selected according to the similarity to a user's query. This approach reduces hallucinations, especially when handling knowledge that is underrepresented or absent in the pretraining data, such as long-tail or new knowledge.

Several recent benchmarks highlight the critical role of RAG in the management of long-tail entities and relations. Mallen et al. (2023) introduce PopQA, a Wikidata-derived QA dataset with controlled entity popularity, showing that while larger models perform poorly on rare facts, RAG improve performances for low-popularity queries. Similarly, Sun et al. (2024) propose Head-to-Tail, a DBpedia-based benchmark that documents a sharp performance drop in infrequent entities when LLMs rely solely on their parametric memory. Maekawa et al. (2024) presents WiTQA, which stratifies Wikidata questions by entity and relation frequency, demonstrating that adaptive retrieval outperforms both pure generation and always-on RAG.

More recently, MINTQA He et al. (2025a) has addressed the dual challenges of unpopular and newly emerging knowledge by posing complex, multi-hop questions. It draws triples from English Wikidata and uses GPT-4o to generate questions spanning one to four hops. MINTQA consists of two subdatasets: MINTQA-POP targets unpopular versus popular facts, and MINTQA-TI contrasts new versus old knowledge. The authors find that as hops increase, the effectiveness of a RAG approach drops.

Although RAG has proven its value for QA, it has yet to be applied to structured KE. Our pipeline fills this gap by proposing to use RAG for long-tail KG enrichment: given a NIL entity as the subject, we retrieve relevant passages from specialized historical corpora and craft prompts that guide an LLM to predict the values of core Wikidata properties, effectively generating RDF triples ready for integration into Wikidata.

2.4 Knowledge Graph Construction and Completion from text

Deep learning has driven rapid advances in automatic KG construction from text. This is reflected in dedicated workshops such as "Deep Learning for Knowledge Graphs" Alam et al. (2023) and "Deep Learning meets Ontologies and Natural Language Processing" Groth et al. (2022), which showcase diverse neural approaches to the problem. In particular, Text-to-AMR (Abstract Meaning Representation) transduction has emerged as a versatile paradigm, founding KG construction on a textual semantic parsing task. Neural sequence-to-sequence parsers, ranging from SPRING and transition-based AMR parsers Zhou et al. (2021) to multilingual and multiformalism models like XL-AMR Blloshmi et al. (2020) and SGL Procopio et al. (2021), have been proposed. A recent survey by Wein and Opitz (2024) reviews these Text-to-AMR methods, highlighting their utility in a wide range of engineering tasks, thanks to their ability to capture deep semantic structure. More recently, the rise of transformers and LLMs has reshaped KG completion and enrichment. Yao et al. (2019) pioneered KG-BERT, which frames each triple as a textual sequence and applies BERT to predict missing links. Generative information extraction surveys such as Zhong et al. (2023) document how LLMs can convert unstructured text into structured representations, while Pan et al. (2024) explore the synergy between KGs and LLMs, using KGs to inform the LLM outputs and vice versa. In contrast to purely neural or transformer-only pipelines, Text2AMR2FRED Gangemi et al. (2023); Gangemi and Nuzzolese (2025) retains formal semantics end-to-end by using AMR graphs as an intermediate layer and then mapping them into OWL-compliant RDF aligned with PropBank, WordNet,

DBpedia, and DOLCE. This alignment supports ontology-level reasoning, semantic querying, and integration with established vocabularies, features that are often overlooked in LLM-centric approaches. Text2AMR2FRED has been successfully applied to diachronic corpora in the cultural heritage domain Gangemi et al. (2024).

Unlike text-to-KG pipelines, which extract triples from a given text span, our pipeline retrieves context from multiple passages and maps values to core properties, producing entity-wise KG modules that can be ingested into Wikidata.

Several recent frameworks target KG completion by harvesting facts from web pages or other semi-structured/unstructured sources. Guo et al. (2023) enriches Wikidata entities by crawling their linked web pages, casting property filling as an extractive QA task: for each missing property, they generate a question from the property's semantic label, apply a fine-tuned QA model to extract candidate values, and resolve them to KB entries via a custom linking module. They are based on Levy et al. (2017), which frames the KB population as relation extraction through reading comprehension questions, and Kratzwald et al. (2020), which completes KBs using a QA pipeline over Wikipedia.

Unlike these systems, which enrich only existing Wikidata entries and rely on entity-to-URL mappings to retrieve web documents, our approach retrieves context from specialized historical corpora and generates RDF triples for entirely absent (NIL) entities, enabling Wikidata enrichment with long-tail historical figures.

2.5 Bias in Wikidata

Search engines, LLMs, and KGs differ in the extent of knowledge they represent: search engines reflect the distribution of content on the Web; LLMs are trained on large-scale textual data and may overlook long-tail facts; KGs like Wikidata offer structured and curated representations, but with limited scope. As noted by Hogan et al. (2025), even the largest KGs capture only a small fraction of what appears in web-scale corpora. These coverage gaps affect the ability of data-driven systems to meet information needs and can influence downstream behavior Abián et al. (2022). Building on these observations, recent studies have advanced the systematic examination of content gaps in Wikidata. Ripoll et al. (2025) provides a comprehensive review of the typologies of gaps, including demographic, socioeconomic, occupational, and geographic dimensions. Particularly relevant to our work is their discussion of recency bias, namely the tendency of knowledge bases to favor contemporary subjects over historical ones. This bias intersects with gender disparities: for instance, the ratio of Wikipedia articles on women to men remains low for individuals born before the 19th century but increases thereafter, with projections indicating parity for those born in 2034.

Laouenan et al. (2022) documents gender bias patterns in their cross-verified database of 2.29 million notable individuals spanning 3500 BC to 2018 AD. The percentage of female individuals exhibits a U-shaped pattern over time, reaching a minimum of 5–10% around 1750 before rising to 20–30% by 2018. This increase correlates with the growing prominence of the Sports and Culture categories, which show more gender balance than the Leadership and Academic fields. The study finds that Western non-English Wikipedia editions reduce gender bias compared to English Wikipedia alone, achieving 27% female representation versus 22.5% in the English edition after 1950. Yu et al. (2025) analyze gender disparities across 1 million Wikipedia biographies from 2010 to 2020. While women and non-binary individuals achieved parity in cross-language presence by 2015, their articles remain consistently shorter and show lower consensus across language editions compared to men's biographies. This pattern suggests that quantitative representation improvements have not resolved disparities in content depth and consistency.

Despite increasing awareness of biases in Wikidata, existing efforts have largely focused on quantifying gaps or improving representation for entities already present in the graph. However, historical figures, particularly those underrepresented or absent from mainstream reference sources, remain difficult to recover through traditional KG enrichment methods. Our work addresses this limitation by targeting NIL entities from historical corpora and generating structured knowledge about them. In doing so, we not only complement existing approaches to bias mitigation, but also provide a mechanism to counteract recency bias by surfacing long-tail historical individuals and their properties for inclusion in Wikidata.

3 How to automatically enrich Knowledge Graphs with Long-tail Entities

Algorithm 1 Automatic enrichment of a NIL entity

Require: A sentence s mentioning the NIL entity e of Wikidata type T
Require: $\mathcal{C}$ a corpus that possibly contains information on e
Require: A set of retrievers $\mathcal{R}$, a set of LLMs $\mathcal{LLM}$ and a confidence threshold τ

1: $KG_e \leftarrow \emptyset$
2: **for all** $P \in$ *top-k* properties for T **do**
3: $T_P \leftarrow \emptyset$
4: **for all** $(R, L) \in \mathcal{R} \times \mathcal{LLM}$ **do** ▷ All combinations of retrievers and LLMs
5: Instantiate the query template q for L using s, e, P
6: Retrieve the context c related to e from $\mathcal{C}$ using query t and retriever R
7: Combine the context c and the query t into the prompt p
8: Prompt L for the answer $\hat{a}$ using p
9: Link the answer $\hat{a}$ to a literal or a Wikidata entity $\tilde{a}$
10: Add $\langle e, P, \tilde{a} \rangle$ to T_P
11: **end for**
12: Add $\langle s, P, majority(T_P) \rangle$ to KG_e
13: **end for**
14: **return** KG_e

Algorithm 1 describes our pipeline using pseudo-code. Given a NIL entity and its type, it initializes an empty KG related to the NIL entity (line 1) and enumerates the top-k properties that will be used to enrich the entity (line

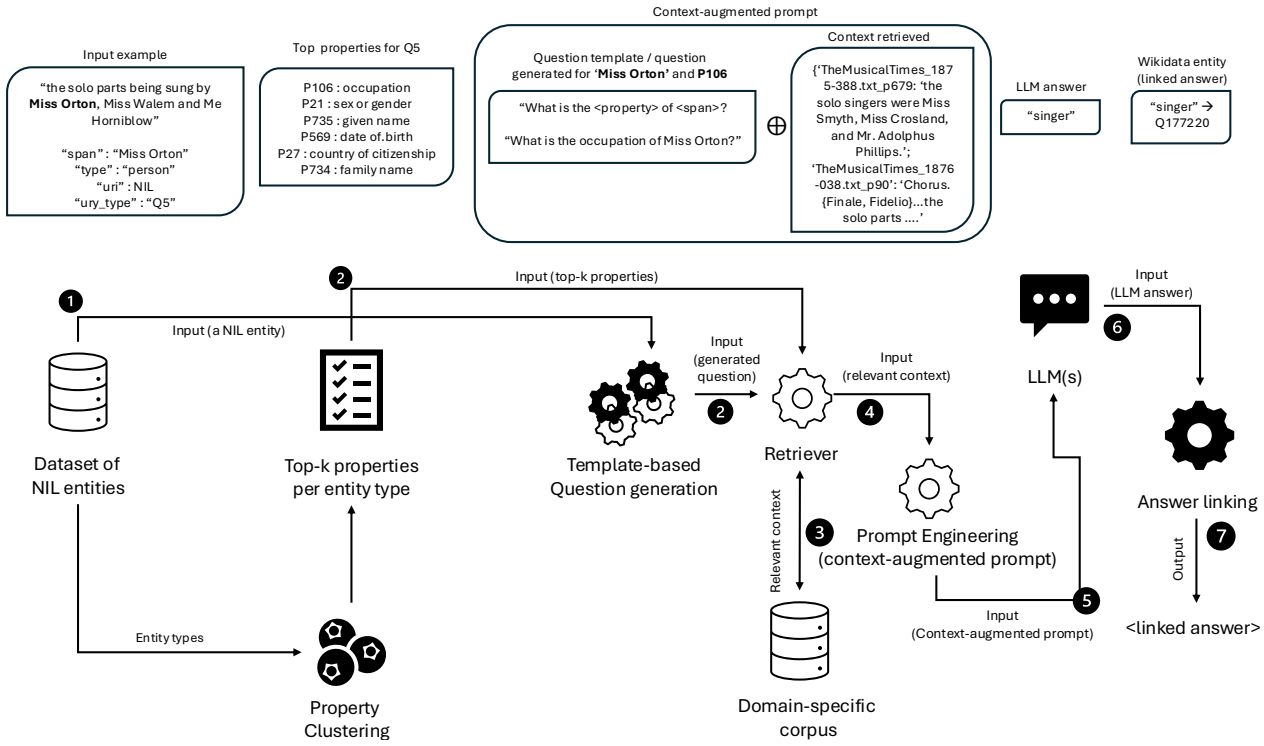


Figure 1. Extracting knowledge about NIL entities to enrich Wikidata: example of data extracted and generated for a relevant historical figure in the music domain, and diagram of the end-to-end pipeline: document processing, question generation, retrieval, answer generation, and answer linking.

2). Although it is possible to manually specify the set of properties, we propose to automatically extract them using the K-means algorithm (Section 3.1). For each combination of LLMs and context retrievers (line 4) we first search the domain-specific corpus for relevant passages given the sentence mentioning the NIL entity and the property at hand (line 5 and 6, Section 3.2). We then rely on this information to instantiate a prompt for the LLM (line 7, Section 3.3) and use it to obtain a answer for the property at hand (line 8, section 3.4). We link the answer to Wikidata QIDs (line 9, Section 3.5) and store the candidate triple (line 10). Once all triples have been generated, multiple combinations of LLMs and retrievers likely provide the same linked answer. We experiment with a majority-voting algorithm to identify the linked answers most frequently predicted (line 12, Section 3.6).

This retrieval-augmented generation approach enables the extraction of property values from corpus evidence and structures them according to Wikidata conventions. In the next sections we provide a detailed description of the different steps of our pipeline. To better illustrate the method we will rely on a running example based on the following sentence:

> the solo parts being sung by Miss Orton, Miss Waleman and Me Horniblow

where we are interested in enriching the entity *Miss Orton* of type *person* (Q5[3]). The running example and a schema of our method are reported in Figure 1.

3.1 Automatically detect most prominent properties

In order to automatically extract a set of properties that are likely to be predicted by an LLM, we rely on a distributional hypothesis. Given a type of entity in Wikidata, for example *person* (Q5), we argue that external corpora and the parametric memory of LLMs are more likely to contain information on the most used properties for persons, for example *occupation* (P106), rather then more specific properties, such as *record label* (P264).

We extract this set of properties by first collecting all the entities of the specified type from Wikidata, and for each property used by at least one entity, count the total number of entities using that property. We then group properties using the K-means algorithm and vary K from 1 cluster to N clusters. We select the final number of clusters by relying on the elbow method: we compute the sum of squared errors (SSE) for each number of clusters and select the point that corresponds to an inflection of the obtained curve.

For example, the we might obtain that the most frequent properties for *person* are *occupation* (P106) and *country of citizenship* (P27). Hence, we will enrich the entity related to *Miss Orton* using those properties.

3.2 Retrieve informative context from the corpus

Once the set of properties are identified, we can prepare the queries for the LLMs. The most critical step is the context retrieval phase, which provides additional information

[3] https://www.wikidata.org/wiki/Q5

relevant to the NIL entity to the LLM. Even though the entity is not available in Wikidata (and hence probably on Wikipedia), we argue that the vast training corpus of LLMs might still contain information on that entity. In this case, even though the retrieved context might not contain direct information about a specific property, it still acts as a hook onto its parametric memory. For instance, in our use case, we expect the LLM to give emphasis to information memorized from documents related to the same historical period as the fragments we provide.

We frame the KE phase as a QA task. We construct the query by concatenating the entity mention, its sentence, and the property-specific question using the fixed prompt:

> Entity mention: `{ mention }`
> Source sentence: `{ sentence }`
> Property question: What is the `{ property }` of `{ entity }`?

The retriever ranks paragraphs from our provided corpus against the query. To avoid an over-abundance of information, we only consider the top-k passages retrieved.

For example, given the property *occupation*, the query posed to the retriever will be

> Entity mention: *"Miss Orton"*
> Source sentence: *"the solo parts being sung by Miss Orton, Miss Walem and Me Hornblow"*
> Property question: *"What is the occupation of Miss Orton?"*

and the retrieved passages might be:

- *"TheMusicalTimes_1873-368.txt_p579: 'the solo singers were [...]"'*
- *"TheMusicalTimes_1876-038.txt_p90: 'Chorus. (Finale, Fidelio)...the solo parts [...]"'*

3.3 Generate the question to the LLM

Given some context related to the entity and property at hand, we instantiate the prompt for the LLM by combining the query used with the context retriever and the retrieved context into a structured prompt containing four components:

- Task instruction specifying the required answer format
- Demonstration example
- Input block containing the entity mention, sentence, retrieved context, and question
- Answer trigger keyword (*Answer:*)

Examples of a prompt is reported in Appendix .1.

3.4 Prompt the LLM for an answer

We prompt the LLM to answer the question and obtain an answer for the property at hand using the prompt constructed in previous section. Note that when relying on multiple retrievers, an LLM will be prompted multiple times with similar prompts, where the main differences are in the contextual information provided. This will step can produce different *candidate* answers, depending on the effectiveness and informativeness of the context provided.

For example, the LLM may generate the textual answer*"singer"* for the prompt related to the example sentence from previous section. Note that the candidate answer is in free-form, it is not constrained to be a valid Wikidata entity. This introduces two main issues. First, the answer cannot be directly used within Wikidata. For instance, the *singer* occupation is represented as an entity in Wikidata (`Q177220`). Secondly, when producing different candidate answers using different retrievers and LLMs, the answers might be not comparable due to superficial differences. For instance, *singer* and *vocalist* are both valid candidate answers that refer to the same Wikidata entity.

3.5 Link the answers to Wikidata

We overcome the ambiguity in free-form answers by automatically aligning free text candidate answers to Wikidata QIDs, effectively generating valid Wikidata triples. This is done by building a *lookup table* for each property considered, as shown in pseudocode in Algorithm 2.

Algorithm 2 Creation of lookup table for a property P for entities of type T

Require: A Wikidata type T and a Wikidata property P
$\mathcal{LT}_P \leftarrow \emptyset$ ▷ The lookup table
for all $\langle s, p, o\rangle \in$ Wikidata s.t. $type(s) = T$ **do** ▷ $\langle s, P31, T\rangle \in$ Wikidata
  for all l **such that** l is a label or an alias of $type(o)$ **do**
    $\mathcal{LT}_P[l] = type(o)$
  end for
end for

Intuitively, given a property P and the type T of the NIL entity considered, we collect all the Wikidata entities of type T for which a triple $\langle e, P, o\rangle$ exists with $type(e) = T$. We then collect all labels and aliases of $type(o)$ and store their association to $type(o)$ in the lookup table of P. Intuitively, the lookup table contains superficial mentions of the values that can be taken by property P and maps them to their corresponding Wikidata entry.

The lookup table can be used to disambiguate the answer of an LLM to a Wikidata entry. We employ a multistage string matching approach to accommodate variations in entity mentions. If a direct lookup on the table fails, we transform the LLM answer by normalizing case differences, diacritical marks (e.g., "José" versus "Jose"), and punctuation differences. If an answer contains multiple components separated by some delimiter (e.g., "composer; conductor"), we process each component independently, producing multiple valid values for that property. If the matching process fails, we discard the candidate LLM answer, reducing spurious matches.

For example, for *"Miss Orton"*, the LLM candidate answers *"singer"* and *"vocalist"* are both linked to `Q177220`, the QID of the entity *singer*. It is hence possible to generate valid Wikidata triples from both candidate answers.

3.6 Majority voting

Each answer is the result of a pipeline involving the entity, its context, a retrival method, and an LLM. It is reasonable to assume that, since the target NIL entity is the same for each entity, most of the candidate answers will converge towards a unique entity, while some combinations will provide outliers that should be ignored. Here, we hypothesize that we can improve the accuracy of the results by combining multiple answers and applying a majority vote.

To identify this entity, we rely on the Boyer-Moore majority voting algorithm, which efficiently identifies the answer repeated >50% of the times. For example, considering the property "country of citizenship", given the candidate answers `France (Q142)`, `France (Q142)`, `Germany (Q183)`, `United Kingdom (Q145)`, `France (Q142)`, the algorithm will identify the entity `France (Q142)` as the final answer. For properties accepting multiple values, the algorithm will accept all the entities repeated >50% of the times.

4 Experiments on Historical Figures

We experiment with the method proposed in Section 3 by adopting the *Musical Heritage Named Entities Recognition, Classification and Linking* (MHERCL) dataset Graciotti et al. (2025a), a component of KE-MHISTO Graciotti et al. (2025b). MHERCL builds upon the Polifonia Corpus Polifonia Corpus and contains sentences from British periodicals published between 1823 and 1900. Those sentences have been extracted from the *Periodicals* module, which includes issues of nineteenth-century music periodicals including *The Harmonicon* and *Dwight's Journal of Music*. Each sentence is annotated for NER (following the Abstract Meaning Representation (AMR) guidelines[4] for consistent entity type classification) and EL tasks, with entities linked to Wikidata where possible, with an inter-annotator agreement of 0.82 Krippendorff's alpha, indicating reliable annotation. Moreover, each sentence includes metadata identifying its original source documents. The dataset contains a total of 2,370 entity mentions (1,805 unique) across 58 different entity types. The five most frequent types are PERSON (1,253 mentions), CITY (262), MUSIC (187), ORGANIZATION (93), and WORK-OF-ART (85).

Entities of type person cover 53% of all annotations and 54% of NIL cases, as can be seen in Table 1. Furthermore, persons are central in KGs, yet their distribution is affected by structural biases such as gender bias. For example, historical figures found in specialised corpora are more likely to be absent from Wikidata Graciotti et al. (2025b), and this deteriorates LLMs performance in downstream tasks Graciotti et al. (2024).

Table 2. Gender distribution: all person mentions versus NIL person mentions in MHERCL.

	Male (%)	Female (%)
All Person NEs	85	15
NIL Person NEs	27	55

Table 1. Distribution of unique NIL entities per type in MHERCL and distributions of person mentions and their gender.

Entity Type	#
person	402
music	118
organization	48
work-of-art	34
opera	21
publication	20
building	18
book	16
worship-place	15
road	9
city	7
company	7
event	4
government-organization	4
festival	3
magazine	2
concert	2
street	2
journal	1
newspaper	1
university	1
school	1
college	1
city-district	1
facility	1
thing	1
scholarship	1
Total person mentions **2370** (85% are Male, 15% are Female) **Total NIL person mentions** **745** (27% NIL are Male, 55% NIL are Female)	

Table 3. Properties of entity of type human (Q5) selected with k-means for person NIL entity KG creation.

QID	Label	#
P106	occupation	11,859,955
P21	sex or gender	9,794,936
P735	given name	8,057,631
P569	date of birth	7,045,476
P27	country of citizenship	5,536,211
P734	family name	5,495,303

Moreover, MHERCL reveals pronounced gender imbalance patterns (Table 1). Male figures dominate overall person mentions (85%) but constitute only 27% of NIL entities. Conversely, women represent 15% of total mentions yet comprise 55% of missing persons. This inversion indicates that women's contributions in nineteenth-century periodicals remain underrepresented in Wikidata. By targeting these NIL entities we address both coverage and representational gaps. Finally, entities of type person own an heterogeneous set of properties on Wikidata, with different data types and cardinality, which allows testing our method on different type of properties.

[4] https://github.com/amrisi/amr-guidelines/blob/master/amr.md#named-entities

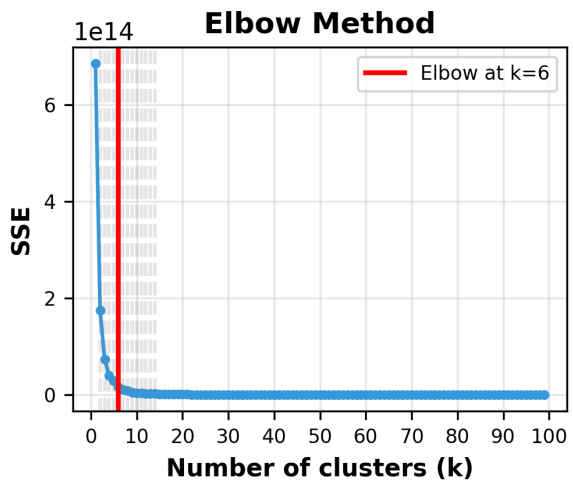


Figure 2. Elbow curve showing the SSE for K-means clustering with $k = 1$ to $k = 100$. The inflection point suggests an optimal grouping at $k = 6$.

In the next sections we give a detailed description on the identification of properties to extract for NIL entities (Section 4.1), the extraction of person entities from MHERCL and our experiments with a wide range of state-of-the-art LLMs and context retrievers. In particular, in Section 4.2, we present i QID-KG, a KG constructed by extracting information on the entities of type person from MHERCL; ii NIL-KG, a **novel manually curated** KG where we enrich some of the NIL entities found in MHERCL with information sourced from the Polifonia corpus and from the Web. Next, we experiment with our method on QID-KG and NIL-KG in Section 4.3.

4.1 Identify the set of properties for NIL persons

As described in Section 3.1, we identify the best set of properties for entities of the given type. In our experiments, we consider the type person and use a complete Wikidata dump for analysis[5]. We first collect all entities that are subject to a triple with property `P31` (`instance of`) and `Q5` (*person*). We then collect the properties used by the selected entities and count their number of occurrences. We cluster the properties into different numbers of clusters, as explained in Section 3.1, varying the number of clusters from 1 to 100.

We obtain the eblow curve in Figure 2, which displays a clear inflection at 6 clusters. We hence select the six most frequent properties for extraction, excluding the property `P31` (instance of) which is trivial and is always equal to `Q5`. The final set of core properties selected is reported in Table 3.

4.2 Introducing the QID-KG and NIL-KG Knowledge Graphs

In order to evaluate LLMs performances on predicting NIL entities' attributes, we manually construct two reference KGs using the set of properties of Table 3. We propose the *QID-KG*, which includes MHERCL persons that have a corresponding Wikidata entity, and *NIL-KG* for persons that do not have a Wikidata entity.

Table 4. Property coverage in gold-standard KGs. Numbers in parentheses show percentages.

Property	QID-KG (n=492)	NIL-KG (n=113)
Country of citizenship	416 (84.6%)	87 (77.7%)
Date of birth	479 (97.4%)	17 (15.2%)
Family name	399 (81.1%)	98 (86.7%)
Given name	484 (98.4%)	57 (50.9%)
Occupation	485 (98.6%)	105 (93.8%)
Sex or gender	491 (99.8%)	101 (90.2%)

Table 5. Gold-standard triples for NIL entity *Agnes Larkcom*, extracted from the corpus document *"Highly commended — Annie Albu, Amy Aylward, Jessie Jones, Agnes Larkcom, and Marian Williams."*(*The Musical Times*, 1876) and a web page as additional evidence[7].

Property (ID)	Object (QID)
given name (P735)	Agnes (Q394431)
sex or gender (P21)	female (Q6581072)
country of citizenship (P27)	United Kingdom (Q145)
occupation (P106)	soprano singer (Q98834068), voice teacher (Q7939609)
date of birth (P569)	1865
family name (P734)	NIL

Table 6. Gender distribution in gold-standard KGs.

Gender	QID-KG (492)	NIL-KG (113)
Male	439 (89.2%)	64 (57.1%)
Female	52 (10.6%)	37 (33.0%)
Missing	1 (0.2%)	11 (9.8%)

For each non-NIL person entity in MHERCL we query Wikidata and retrieve the values of each property. For each entity we store the QID or literal value of each property[6], resulting in the QID-KG Knowledge Graph. Table 4 shows property coverage ranging from 84.6% (country of citizenship) to 99.8% (sex or gender).

On the other hand, we manually annotate the properties of 113 NIL entities in MHERCL examining both source periodicals and reliable Web resources.

Table 5 shows the annotations and their source within the corpus and other web sources for the NIL entity *Agnes Larkcom*. The final set of annotation is used to create the NIL-KG Knowledge Graph, which includes 465 manually curated triples. We find property coverage to significantly vary based on the entity. For instance, we identify the occupation for 93.8% of the entities but we only find a reliable date of birth for 15.2% of the entities. Table 4 reports the coverage for each property in NIL-KG.

Since QID-KG directly leverages the entities annotated in MHERCL, we find that it similarly exhibits a pronounced male dominance with minimal female representation, as reported in Table 6. NIL-KG exhibits a less pronounced dominance with respect to QID-KG, even though female entities are still under-represented. This shift confirms that

[5] The Wikidata dump used for this analysis was downloaded on 29/03/2025.
[6] Dates are stored as `xsd:date` literals; all other values as QIDs.
[7] https://shigovoicelessons.com/voicetalk/2016/04/the-garcia-lineage-agnes-larkcom.html.

historical women face double marginalization: underrepresented in period sources and overlooked in contemporary knowledge bases.

4.3 Experiments on QID-KG and NIL-KG

We experiment with the method of Section 3 on both QID-KG and NIL-KG, introduced in the previous section. For each person in both KGs we retrieve the information of the properties identified in Table 3. All context is retrieved by querying the *Periodicals* module in the Polifonia corpus.

Table 7. Retrieval methods.

Type	Model
Sparse	BM25 Robertson and Zaragoza (2009)
Dense	`contriever` Izacard et al. (2022) `gtr-t5-large` Ni et al. (2022) `gtr-t5-xxl` Ni et al. (2022) `bge-large-en` Xiao et al. (2024) `instructor-xl` Su et al. (2023)

Table 8. LLMs used.

Parameters	LLM
7-14B	Mixtral-8×7B Jiang et al. (2023) Phi-3-Medium Abdin et al. (2024)
27B	Gemma-2-27B-IT et al. (2024)
70-72B	Llama-3.3-70B Grattafiori et al. (2024) Qwen-2.5-72B Hui et al. (2024)
est. ∼200B+	GPT-4o-mini

We experiment with six different information retrieval methods based on sparse and dense retrieval techniques, listed in Table 7, and with six different LLMs across different parameter scales, listed in Table 8 and selected following the comparative study of He et al. (2025b). We only collect the top 3 documents retrieved by a retriever and query LLMs using the OpenRouter API[8], which provides a uniform interface across model backends.

When querying an LLM, we experiment with two variants of the prompt from Section 3.4, a **RAG** prompt where we provide the context retrieved and a **zero-shot** prompt where we do not provide any context. This allows us to measure the impact of the retrieved context.

In total, we produce 72 candidate answers for each property when combining zero-shot and RAG results.

During the majority voting phase described in Section 3.6 we enforce a strict majority, meaning that a candidate answer is selected as the final value only if it covers *at least* half of the set of answers. We perform majority voting for each LLM separately, to assess the accuracy of the LLM, and by combining all the candidate answers from all the LLMs.

4.3.1 Evaluation Metrics We evaluate our experiments for each property using standard information retrieval metrics on the linked answer (Section 3.5). We filter out candiate answers longer than 100 characters.

Table 9 shows examples of our evaluation setting. Each prediction is classified as True Positive (TP) when the system correctly identifies a property value matching the gold standard, False Positive (FP) when the system produces an incorrect property value, False Negative (FN) when the system fails to identify an existing property value, and True Negative (TN) when the system correctly identifies that no property value exists.

Table 9. Examples of evaluation metrics.

Target	Pred	Result
Q145	Q145	TP
Q145	Q30	FP ∧ FN
Q145	NIL	FN
NIL	NIL	TN
NIL	Q145	FP

For properties that allow multiple correct values, such as *occupation*, we evaluate whether there is a partial match between the gold standard and the linked answers (binary classification: at least one correct answer has been predicted) as well as the completeness of the answers (multilabel classification). We heuristically extract answers by splitting on commas, semicolons, and pipe characters.

4.3.2 Evaluation on dates For the Date of Birth (DoB) property, we assess performances at four levels:

1. exact match on the date;
2. matching year, e.g. *1810-03-15* and *1810-12-01*;
3. matching decade, e.g. *1810-03-15* and *1814-04-12*;
4. matching century, e.g. *1810-03-15* and *1858-03-15*

Dates are normalized from various date formats including ISO dates ("1810-03-15"), dates with time zones ("1810-03-15T00:00:00Z"), textual representations ("born on March 15, 1810"), and year-only specifications. The different criteria on date evaluation capture the ability to extract temporal information at different precision levels, accounting for cases in historical texts where exact dates may be uncertain or differently expressed, and allows us to assess whether added context enforces this aspect.

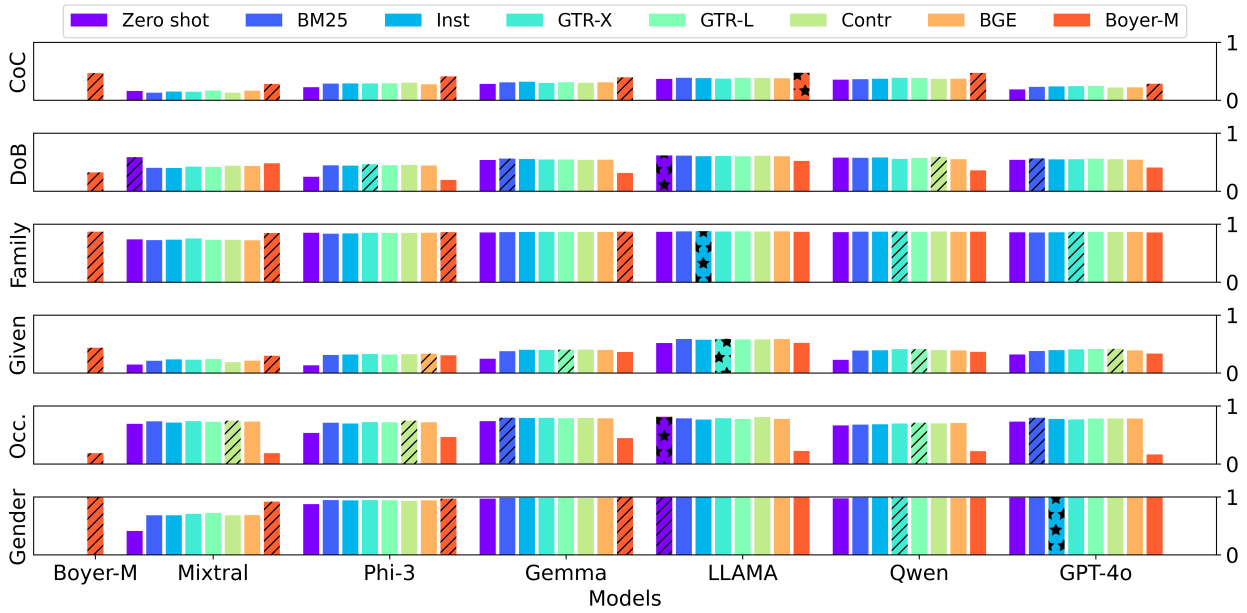
5 Results

This section reports the experimental results of our proposed KE pipeline on the setting described in previous section. We first present the performance of the methods for each property on both **QID-KG** and **NIL-KG**. Then, we provide granular evaluations specific to dates and occupations and examine the contribution of retrieval and majority voting.

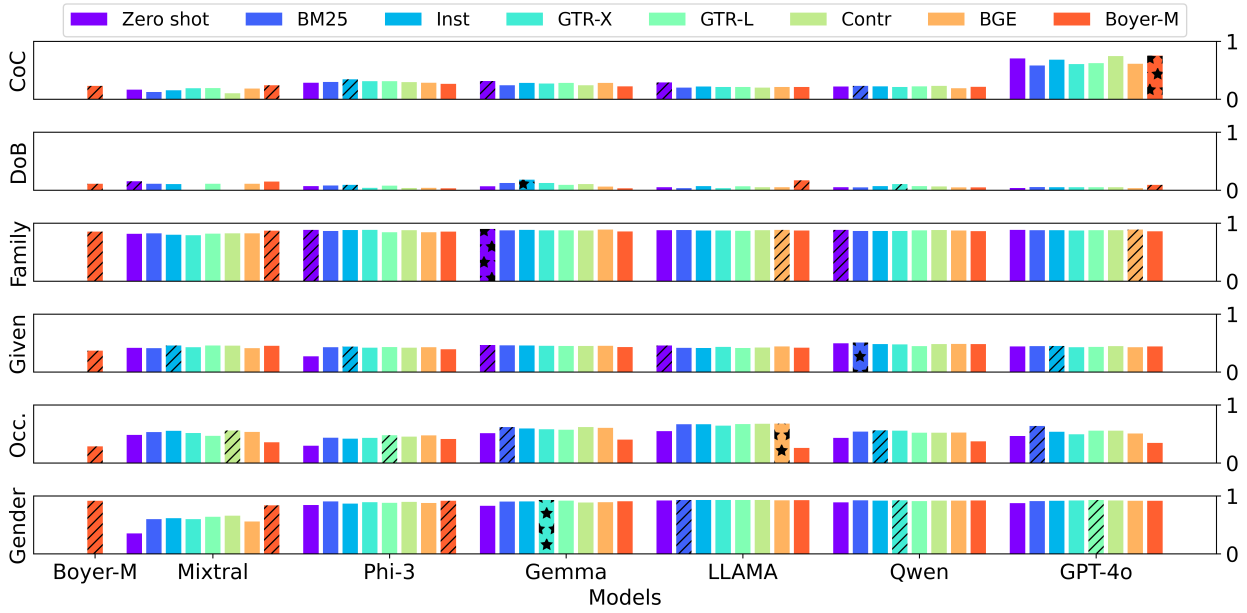
5.1 Overall Property Extraction

Figure 3 reports the F1 score for each property of Table 3 on QID-KG (Figure 3a) and NIL-KG (Figure 3b). We report complete evaluation tables with P, R, and F1 for each property in the Appendix .2. Experiments on QID-KG scores systematically exceed NIL-KG scores, reflecting the challenge of extracting properties for entities absent from Wikidata. Without pre-existing records, LLMs cannot leverage parametric memory.

[8] https://openrouter.ai/



(a) QID-KG



(b) NIL-KG

Figure 3. F1 measures computed on QID-KG and NIL-KG for each property. Best results for each model are represented by hatched rectangles. The best overall result is filled with star symbols.

Minimal context suffices for family name and sex or gender. `Family name` provides an exception to higher QID-KG scores. Performances differ marginally between KGs since surnames appear in the entity mention itself, requiring minimal external evidence. `Sex or gender` achieves the highest performance across both KGs, with `family name` second. Sex or gender extraction relies on local cues, such as honorifics and pronouns, enabling consistent high performances.

Majority voting improves CoC performance on QID-KG. For `Country of Citizenship`, out majority voting strategy combining all the different candidate answers achieves the best performances, $F1 = 0.47$. Moreover, relying on the majority vote for each single model is consistently the best approach, indicating that exploiting different candidate answers reduces the chance of noisy predictions. Nonetheless, when tested on NIL-KG, single-retriever approaches often match or exceed combining candidate answers. We find GPT-4o using majority voting to be the best performing approach on NIL-KG (F1 = 0.75).

LLM priors dominate for DoB; retrieval offers limited help. For `date of birth`, majority voting hampers performance when compared to the best individual configurations on QID-KG. Results are mixed on NIL-KG: majority voting typically increases performance, except for Phi-3 and Gemma. Zero-shot configurations match or exceed retrieval-augmented runs in both KGs (except Phi-3 on QID-KG and Gemma on NIL-KG). This reflects

data sparsity: nineteenth-century periodicals focus on contemporary events, concerts, performances, or reviews that rarely mention birth dates. Hence, LLMs must rely on parametric knowledge for known figures' dates and exploit the context to grasp the historical period. The best performances are obtained using Llama-3.3-70B on a zero-shot setting with QID-KG (0.62), and Gemma-2-27B-IT with the instructor-xl retriever on NIL-KG (0.18). Date of birth shows the lowest performance on NIL-KG.

No consistent gains for given name from RAG or voting. For `given name`, results vary across configurations. Majority voting and RAG do not provide any systematic improvement over zero-shot approaches, both on QID-KG and NIL-KG. The best performances are obtained using Llama-3.3-70B with the BM25 retriever (or GTR-X retriever) on QID-KG (F1 = 0.59), and using Qwen-2.5-72B with the BM25 retriever on NIL-KG (F1 = 0.51).

Retrieval boosts occupation accuracy, voting hurts recall. For the `occupation` property, the retrieval-augmented configuration yields consistent gains. On the QID-KG, every model improves its performance when context passages are used, with the largest improvements observed on smaller models, like Mixtral and Phi-3. The effect is more prominent on the NIL-KG, where context-aware runs consistently outperform the zero-shot baselines, confirming that external evidence is essential when the target entity is absent from Wikidata. Notably, every model (except LLAMA on QID-KG, which performs better on a zero-shot setting) shows more prominent improvements on NIL-KG than on QID-KG, underscoring the increased utility of retrieval when parametric memory alone is insufficient. A majority voting approach to the `occupation` property leads to lower performances on both KGs. This is because the strict majority policy discards many candidate answers, which result in a lower possible recall.

5.2 Date of Birth Evaluation

We evaluate the extraction of `date of birth (DoB)` using four different criteria, as explained in previous section, and report its results in Figure 4 for both QID-KG (4a) and NIL-KG (4b). We report complete evaluation tables with P, R, and F1 for each property in the Appendix .3.

Overall, performances increase when a coarser evaluation criteria is used across all models and retrievers. This confirms the tendency of LLMs to provide temporally approximate answers, even when the precise date is unavailable. We remark that F1 scores are consistently higher on the **QID-KG** than on the **NIL-KG**, regardless of the criteria. This disparity reflects the fundamental challenge posed by NIL entities: many have no date of birth attested in any structured or unstructured source, so an LLM should be able to reply that no DoB is known for that entity. Instead, LLMs usually return an answer, even if it is wrong. In contrast, entities in the QID-KG often appear in widely available historical databases, enabling LLMs to rely on parametric memory and leverage supporting evidence seen during pre-training.

5.3 Occupation Evaluation

Finally we assess the extraction of the `occupation` property as a multi-label classification task, where each entity may have multiple occupation labels. For evaluation, we treat each entity as a multi-label classification instance and compute precision, recall, and F1 per sample. We report its results in Figure 5 for both QID-KG (5a) and NIL-KG (5b). We report complete evaluation tables with P, R, and F1 for each property in the Appendix .3.

Consistent with Section 5.1, retrieval-augmented prompting markedly improves performance. The majority voting approach performs poorly across both KGs while retrieval-augmented approaches show substantial gains: Phi-3-Medium improves its F1 score on QID-KG while on NIL-KG, the improvement is even more pronounced, particularly for GPT-4o.

Llama-3.3 exhibits a distinct precision-recall trade-off: while it has a moderate precision score, it achieves the highest recall across different configurations, significantly outperforming all the other methods on NIL-KG. This contrasts with Gemma-2-27B-IT and GPT-4o, which achieve higher precision but lower recall. This suggests that different LLMs balance differently comprehensiveness over precision when extracting multiple values for a property, in this case occupation.

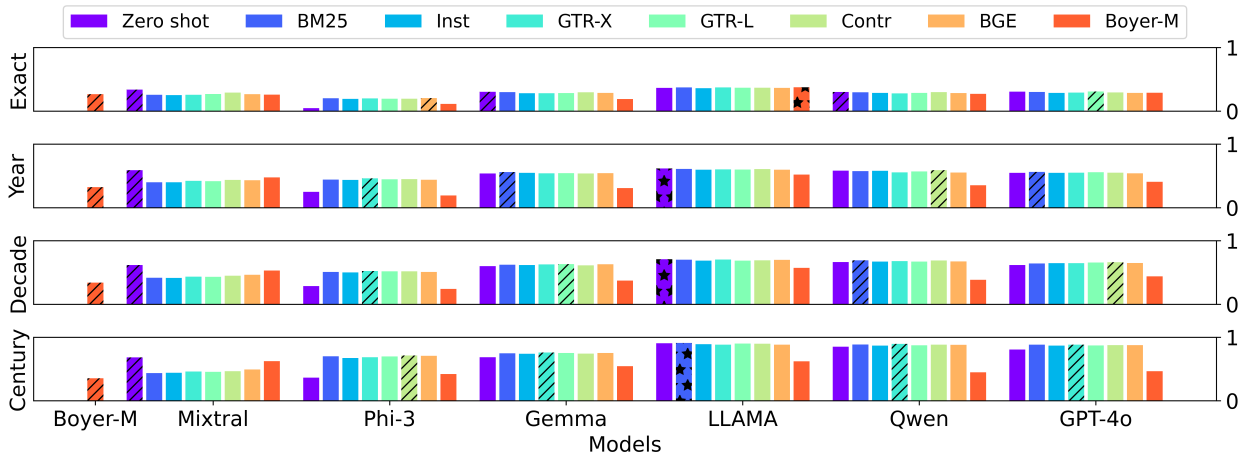
6 Conclusion, Discussion and Social Impact

6.1 Leveraging Historical Sources for Wikidata Enrichment

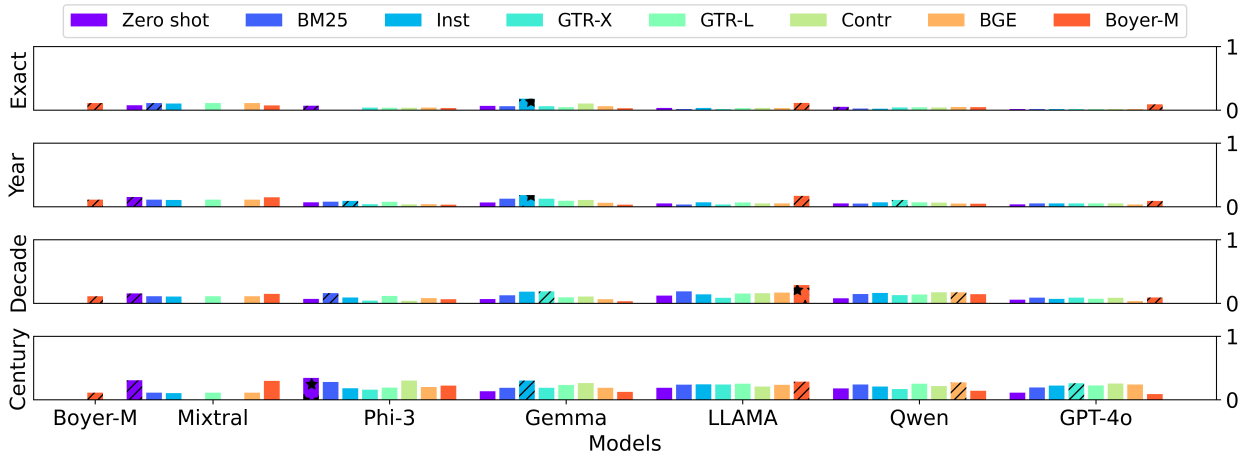
Our results propose a pipeline that leverages historical document corpora to address content gaps in KGs. We show that retrieval-augmented generative AI applied to domain-specific diachronic corpora can construct KGs suitable for Wikidata enrichment, though significant challenges remain.

Pipeline efficacy varies dramatically across properties. Testing on both NIL-KG and QID-KG reveals the influence of parametric memory on performance. For occupation extraction — arguably the most property we consider — our pipeline achieves an F1 score of 0.68 on NIL-KG and 0.81 on QID-KG. This performance gap confirms that, even in low-entity entities, the effect of the pre-training corpora on the parametric memory of an LLM has a substantial influence on its knowledge. Nonetheless, our results provide evidence that it is possible to probe them for knowledge that is not already part of the main knowledge bases, Wikidata in our case.

Occupation extraction particularly demonstrates the influence of retrieval augmentation. On QID-KG, retrieval only slightly improves the F1 score (+0.05), however, a more substantial gain is observed on NIL-KG (+0.1). This confirms our initial hypothesis: retrieval compensates for entities that are underrepresented in the models' parametric knowledge. This suggests that property selection should align with the characteristics of the corpus using during retrieval. We observe better performances on the occupation property because the nineteenth-century periodicals document we use document professional activities. On the other hand, the date of birth fails since they contain few and sparse biographical details.



(a) QID-KG



(b) NIL-KG

Figure 4. F1 score computed on QID-KG and NIL-KG for the date of birth using Exact, Decade and Century criteria. Best results for each model are represented by hatched rectangles. The best overall result is filled with star symbols.

6.2 Model-Retriever Configuration Analysis

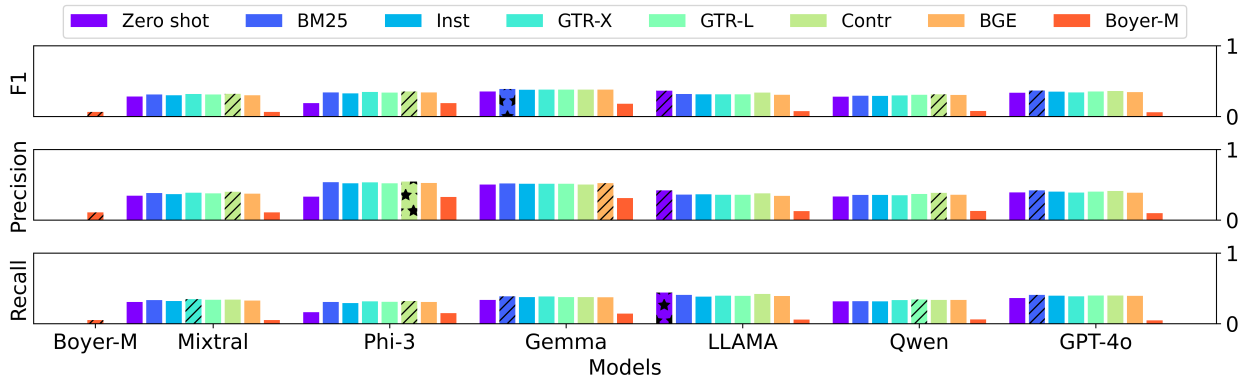
We do not identify a single model-retriever combination that dominates performance across all properties. Llama-3.3-70B achieves the highest performance on date of birth solely relying on its parametric knowledge (0.62 F1 on QID-KG in a zero-shot setting). Gemma-2-27B-IT with dense retrievers excels at multi-label occupation classification (0.40 median F1 on both KGs). GPT-4o paired with Contriever achieves 0.75 F1 for country of citizenship on NIL-KG, yet performs poorly on QID-KG for the same property.

This variability reflects three factors: (1) property-specific extraction challenges, (2) corpus evidence distribution, and (3) model-retriever synergies. Properties that rely on sparse information, like date of birth, benefit from parametric memory while more knowledge-intensive and contextual properties, such as occupation, require retrieval that aggregates meaningful context across different relevant documents. Nonetheless, extracting relevant context is a non-trivial challenge. The performances on the country of citizenship property, which is intuirively simpler than the occupation or date of birth property, show unstable performances, suggesting issues dependent on the extraction quality of the retriever. These patterns require property-specific configuration, rather than universal deployment strategies.

6.3 Parameter Scaling Effects

Surprisingly, we find that increasing model parameters does not guarantee a performance improvement. While Llama-3.3-70B (70B) performs best in date extraction, smaller models match or exceed its performance on other properties. Phi-3-Medium (14B) demonstrates the largest improvement when a retrieval augmentation approach is used: +0.21 F1 for occupation on QID-KG and +0.18 F1 on NIL-KG. This finding suggests that it is feasible to compensate for limited parametric knowledge on smaller models through an effective context-retrieval phase.

Gemma-2-27B-IT (27B) achieves optimal balance, maintaining consistent performance across properties without the computational overhead of bigger models. Mixtral-8×7B shows comparable improvements to Phi-3 when prompted with additional context. These findings challenge the assumption that historical KE requires maximum-scale models. Instead, mid-range models coupled with retrieval augmentation techniques often outperform larger models.



(a) QID-KG

(b) NIL-KG

Figure 5. Precision, Recall and F1 computed on QID-KG and NIL-KG for the occupation property. Best results for each model are represented by hatched rectangles. The best overall result is filled with star symbols.

6.4 Accuracy-Cost Trade-offs

Our results enable property-specific deployment decisions. For sex or gender and family name extraction, zero-shot configurations of smaller models or tailored heuristics are sufficient. Occupation extraction, however, justifies larger models with retrieval, as performance increse on NIL entities. Date of birth presents a complex case: larger models improve year-level accuracy but corpus sparsity limits absolute performance. These patterns suggest a tiered deployment strategy: (1) lightweight models for surface-derivable properties, (2) medium models with retrieval for biographical attributes, and (3) large models only when parametric knowledge proves essential. Future methods can leverage these findings to optimize resource allocation across different historical corpora and entity types.

6.5 Social Impact and Future Directions

This work provides an operational mechanism for addressing documented biases in Wikidata. By extracting structured knowledge about person NIL entities, we enable systematic addition of underrepresented historical figures. The inversion of the gender distribution between all entities (85% male, 15% female) and NIL entities (27% of the males are NIL, 55% of the females are NIL) confirms that current KG construction methods perpetuate historical visibility gaps.

Future work includes extracting properties of entities across whole historical collections, such as the Polifonia Corpus. Key technical challenges include: (1) cross-document entity resolution and clustering for figures mentioned across multiple sources, and (2) multilingual extraction from non-English periodicals. Social challenges require collaboration with historians to validate the extracted knowledge and establish quality thresholds for Wikidata integration. The long-term impact depends on the community adoption of these methods for the systematic reduction of bias in the knowledge infrastructure.

CO2 Emission Related to Experiments

Experiments were conducted using liteLLM[9] via OpenRouter[10], for a cumulative of 180 hours of computation. It was not possible to establish the exact location and hardware type of backend servers. An estimation of 180 hours of computation on an infrastructure of type Azure in the region westeurope has a carbon efficiency of 0.57 $kgCO_2eq/kWh$. We estimated a cumulative of 180 hours of computation performed on hardware of type A100 PCIe 40/80GB (TDP of 250W) – a credible option for the kind of services used. Total emissions are estimated to be 25.65 $kgCO_2eq$ of which 100 percents were directly offset by the cloud provider.

Estimations were conducted using the MachineLearning Impact calculator presented in Lacoste et al. (2019).

[9] https://www.litellm.ai/
[10] https://openrouter.ai/

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme Polifonia, under grants agreement No 101004746 and No 101178821. Nicolas Lazzari is supported by the FAIR – Future Artificial Intelligence Research Foundation as part of the grant agreement MUR n. 341, code PE00000013 CUP 53C22003630006. Enrico Daga is partly supported by the Climate Misinformation Surveillance and Analysis with Multidimensional GIS CHIST-ERA project.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, and Lijuan Wang et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

David Abián, Albert Meroño-Peñuela, and Elena Simperl. An analysis of content gaps versus user needs in the wikidata knowledge graph. In Ulrike Sattler, Aidan Hogan, Maria Keet, Valentina Presutti, João Paulo A. Almeida, Hideaki Takeda, Pierre Monnin, Giuseppe Pirrò, and Claudia d'Amato, editors, *The Semantic Web – ISWC 2022*, pages 354–374, Cham, 2022. Springer International Publishing. ISBN 978-3-031-19433-7.

Mehwish Alam, Davide Buscaldi, Michael Cochez, Francesco Osborne, D Recupero, et al. Preface-proceedings of the 6th workshop on deep learning for knowledge graphs (dlkg2023) co-located with international semantic web conference 2023. In *CEUR WORKSHOP PROCEEDINGS*, volume 3559, pages 1–4. CEUR-WS, 2023.

Abhishek Arora, Emily Silcock, Melissa Dell, and Leander Heldring. Contrastive entity coreference and disambiguation for historical texts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6174–6186, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.355. URL https://aclanthology.org/2024.emnlp-main.355.

Tom Ayoola, Joseph Fisher, and Andrea Pierleoni. Improving entity disambiguation by reasoning over a knowledge base. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2899–2912, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.210. URL https://aclanthology.org/2022.naacl-main.210/.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. XL-AMR: Enabling Cross-Lingual AMR Parsing with Transfer Learning Techniques. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500, Online, November 2020. ACL. doi: 10.18653/v1/2020.emnlp-main.195. URL https://aclanthology.org/2020.emnlp-main.195.

Baptiste Blouin, Cécile Armand, and Christian Henriot. A dataset for named entity recognition and entity linking in Chinese historical newspapers. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 385–394, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.35.

Hang Dong, Jiaoyan Chen, Yuan He, Yinan Liu, and Ian Horrocks. Reveal the unknown: Out-of-knowledge-base mention discovery with entity linking. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 452–462, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3615036. URL https://doi.org/10.1145/3583780.3615036.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. Diachronic Evaluation of NER Systems on Old Newspapers. In Stephanie Dipper, Friedrich Neubarth, and Heike Zinsmeister, editors, *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 97–107, Bochum, Germany, September 19–21 2016. Bochum, Germany, Bochumer Linguistische Arbeitsberichte. URL https://infoscience.epfl.ch/record/221391?v=pdf.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, page 288–310, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58218-0. doi: 10.1007/978-3-030-58219-7_21.

Maud Ehrmann, Matteo Romanello, Antoine Doucet, and Simon Clematide. Introducing the HIPE 2022 Shared Task: Named Entity Recognition and Linking in Multilingual Historical Documents. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg,

and Vinay Setty, editors, *Advances in Information Retrieval*, volume 13186, pages 347–354, Cham, 2022. Springer International Publishing. ISBN 978-3-030-99738-0 978-3-030-99739-7. doi: 10.1007/978-3-030-99739-7_44.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2), sep 2023. ISSN 0360-0300. doi: 10.1145/3604931. URL https://doi.org/10.1145/3604931.

Gemma Team et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671470. URL https://doi.org/10.1145/3637528.3671470.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0). Web Download, June 2013. Available at http://lemurproject.org/clueweb09/ and http://lemurproject.org/clueweb12/.

Aldo Gangemi and Andrea Giovanni Nuzzolese. Logic augmented generation. *Journal of Web Semantics*, 85:100859, 2025. ISSN 1570-8268. doi: https://doi.org/10.1016/j.websem.2024.100859.

Aldo Gangemi, Arianna Graciotti, Antonello Meloni, Andrea Nuzzolese, Valentina Presutti, Diego Reforgiato Recupero, Alessandro Russo, and Rocco Tripodi. Text2AMR2FRED, a tool for transforming text into RDF/OWL Knowledge Graphs via Abstract Meaning Representation. In *22nd International Semantic Web Conference*, Athens, Greece, November 2023. CEUR Workshop Proceedings. URL https://ceur-ws.org/Vol-3632/ISWC2023_paper_394.pdf.

Aldo Gangemi, Arianna Graciotti, Eleonora Marzi, Antonello Meloni, Andrea Nuzzolese, Valentina Presutti, Diego Reforgiato Recupero, Alessandro Russo, and Rocco Tripodi. MusicBO, an application of Text2AMR2FRED to the Musical Heritage domain. In *20th Extended Semantic Web Conference*, Crete, Greece, May 2024. CEUR Workshop Proceedings.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

Arianna Graciotti, Valentina Presutti, and Rocco Tripodi. Latent vs explicit knowledge representation: How ChatGPT answers questions about low-frequency entities. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10172–10185, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.888.

Arianna Graciotti, Nicolas Lazzari, Valentina Presutti, and Rocco Tripodi. Musical heritage historical entity linking. *Artificial Intelligence Review*, 58(5):140, 2025a. ISSN 1573-7462. doi: 10.1007/s10462-024-11102-9.

Arianna Graciotti, Leonardo Piano, Nicolas Lazzari, Enrico Daga, Rocco Tripodi, Valentina Presutti, and Livio Pompianu. KE-MHISTO: Towards a multilingual historical knowledge extraction benchmark for addressing the long-tail problem. In *Accepted at: Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria, 2025b. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, and Junteng Jia et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Paul Groth, Anisa Rula, Jodi Schneider, Ilaria Tiddi, Elena Simperl, Panos Alexopoulos, Rinke Hoekstra, Mehwish Alam, Anastasia Dimou, and Minna Tamper. The semantic web: Eswc 2022 satellite events. *Lecture Notes in Computer Science*, 2022.

Kunpeng Guo, Dennis Diefenbach, Antoine Gourru, and Christophe Gravier. Wikidata as a seed for web extraction. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2402–2411, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583236. URL https://doi.org/10.1145/3543507.3583236.

Zhaochen Guo and Denilson Barbosa. Robust Named Entity Disambiguation with Random Walks. *Semantic Web Journal*, 9(4):459–479, January 2018. ISSN 1570-0844. doi: 10.3233/SW-170273. URL https://doi.org/10.3233/SW-170273.

Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G. Moreno, and

Antoine Doucet. A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2328–2334, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463255.

Jie He, Nan Hu, Wanqiu Long, Jiaoyan Chen, and Jeff Z. Pan. Mintqa: A multi-hop question answering benchmark for evaluating llms on new and tail knowledge, 2025a. URL https://arxiv.org/abs/2412.17032.

Jie He, Nan Hu, Wanqiu Long, Jiaoyan Chen, and Jeff Z. Pan. Mintqa: A multi-hop question answering benchmark for evaluating llms on new and tail knowledge, 2025b. URL https://arxiv.org/abs/2412.17032.

Nicolas Heist and Heiko Paulheim. Nastylinker: Nil-aware scalable transformer-based entity linker. In *The Semantic Web: 20th International Conference, ESWC 2023, Hersonissos, Crete, Greece, May 28–June 1, 2023, Proceedings*, page 174–191, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-33454-2. doi: 10.1007/978-3-031-33455-9_11. URL https://doi.org/10.1007/978-3-031-33455-9_11.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1072.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Computing Surveys*, 54(4), July 2021. ISSN 0360-0300. doi: 10.1145/3447772. Place: New York, NY, USA Publisher: Association for Computing Machinery.

Aidan Hogan, Xin Luna Dong, Denny Vrandečić, and Gerhard Weikum. Large language models, knowledge graphs and search engines: A crossroads for answering users' questions, 2025. URL https://arxiv.org/abs/2501.06699.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-coder technical report, 2024. URL https://arxiv.org/abs/2409.12186.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2022. URL https://arxiv.org/abs/2112.09118.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Bernhard Kratzwald, Guo Kunpeng, Stefan Feuerriegel, and Dennis Diefenbach. IntKB: A verifiable interactive framework for knowledge base completion. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5591–5603, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.490. URL https://aclanthology.org/2020.coling-main.490/.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Morgane Laouenan, Palaash Bhargava, Jean-Benoît Eyméoud, Olivier Gergaud, Guillaume Plique, and Etienne Wasmer. A cross-verified database of notable people, 3500bc-2018ad. *Scientific Data*, 9(1):290, 2022. doi: 10.1038/s41597-022-01369-4.

Nicolas Lazzari, Arianna Graciotti, and Valentina Presutti. Constrained information retrieval for long-tail knowledge extraction. In *EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024)*, volume 3967 of *CEUR Workshop Proceedings*, pages 1–18, Amsterdam, The Netherlands, November 2024. CEUR-WS.org. URL https://ceur-ws.org/Vol-3967/X-TAIL-2024_paper_3.pdf.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In Roger Levy and Lucia Specia, editors, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1034. URL https://aclanthology.org/K17-1034/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. Retrieval helps or hurts? a deeper dive into the efficacy of retrieval augmentation to language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5506–5521, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.308. URL https://aclanthology.org/2024.naacl-long.308/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546.

Stefano Menini, Rachele Sprugnoli, and Antonio Uva. "who was pietro badoglio?" towards a QA system for Italian history. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 430–435, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1069/.

Sina Menzel, Hannes Schnaitter, Josefine Zinck, Vivien Petras, Clemens Neudecker, Kai Labusch, Elena Leitner, and Georg Rehm. *Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten*, pages 229–258. De Gruyter Saur, Berlin, Boston, 2021. ISBN 9783110691597. doi: doi:10.1515/9783110691597-012.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.669. URL https://aclanthology.org/2022.emnlp-main.669/.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2038–2048, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657891. URL https://doi.org/10.1145/3626772.3657891.

Polifonia Corpus. https://github.com/polifonia-project/Polifonia-Corpus. URL https://github.com/polifonia-project/Polifonia-Corpus. Accessed February 2025.

Luigi Procopio, Rocco Tripodi, and Roberto Navigli. SGL: Speaking the Graph Languages of Semantic Parsing via Multilingual Translation. In *Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 325–337, Online, June 2021. ACL. doi: 10.18653/v1/2021.naacl-main.30. URL https://aclanthology.org/2021.naacl-main.30.

Marisa Ripoll, Neal Reeves, Anelia Kurteva, Elena Simperl, Albert Meroño Peñuela, and Klaus Diepold. Filling in the blanks? a systematic review and theoretical conceptualisation for measuring wikidata content gaps, 2025. URL https://arxiv.org/abs/2505.16383.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL http://dx.doi.org/10.1561/1500000019.

Cristian Santini, Mary Ann Tan, Oleksandra Bruns, Tabea Tietz, Etienne Posthumus, and Harald Sack. Knowledge extraction for art history: the case of vasari's the lives of the artists. In Adrian Paschke, Georg Rehm, Clemens Neudecker, and Lydia Pintscher, editors, *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022)*, volume 3234 of *CEUR Workshop Proceedings*, Berlin, Germany, September 2022. CEUR-WS.org. URL https://ceur-ws.org/Vol-3234/paper7.pdf.

Cristian Santini, Laura Melosi, and Emanuele Frontoni. Named entity recognition in historical italian: The case of giacomo leopardi's zibaldone. In *EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024)*, volume 3967 of *CEUR Workshop Proceedings*, pages 1–18, Amsterdam, The Netherlands, November 2024. CEUR-WS.org. URL https://ceur-ws.org/Vol-3967/X-TAIL-2024_paper_1.pdf.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13 (3):527–570, 2022.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1102–1121. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.71. URL https://doi.org/10.18653/v1/2023.findings-acl.71.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.18. URL https://aclanthology.org/2024.naacl-long.18.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3025–3035, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531734. URL https://doi.org/10.1145/3477495.3531734.

Shira Wein and Juri Opitz. A survey of AMR applications. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen,

editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6856–6875, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.390. URL https://aclanthology.org/2024.emnlp-main.390.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Xingrun Xing. LM-cocktail: Resilient tuning of language models via model merging. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2474–2488, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.145. URL https://aclanthology.org/2024.findings-acl.145/.

Liang Yao, Chengsheng Mao, and Yuan Luo. Kg-bert: Bert for knowledge graph completion. *ArXiv*, abs/1909.03193, 2019. URL https://api.semanticscholar.org/CorpusID:202539519.

Yulin Yu, Xianglong Li, Tianyi Li, Paramveer S. Dhillon, and Daniel M. Romero. Demographic disparity in wikipedia coverage: a global perspective. *EPJ Data Science*, 14(1):15, 2025. doi: 10.1140/epjds/s13688-025-00530-4.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. A comprehensive survey on automatic knowledge graph construction. *ACM Comput. Surv.*, 56(4), November 2023. ISSN 0360-0300. doi: 10.1145/3618295.

Jiawei Zhou, Tahira Naseem, Ramón Fernandez Astudillo, Young-Suk Lee, Radu Florian, and Salim Roukos. Structure-aware Fine-tuning of Sequence-to-sequence Transformers for Transition-based AMR Parsing. In *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Online and Punta Cana, Dominican Republic, November 2021. ACL. doi: 10.18653/v1/2021.emnlp-main.507. URL https://aclanthology.org/2021.emnlp-main.507.

7 Appendix

.1 Example Prompt

Listing 1 shows the exact prompt sent to the language model for the entity *Gavinies* for the `occupation` property. Block markers (`###TASK`, `###EXAMPLE`, `###INPUT`, `###ANSWER_PREFIX`) partition the prompt into its logical sections.

```
###TASK
Your task is to determine the occupation(s)
    of a given person occurring in a
given sentence based on some provided
    relevant context. Your answer MUST
    consist
solely of a semicolon-separated list of
    occupation(s) (e.g., "occupation1;
occupation2; ...") with no extra text.

###EXAMPLE
NOTE: The example below is for demonstration
     purposes only and should not
influence your answer.

Given the person "Bach" occurring in the
    sentence
"... selections from the works of Bach,
    Handel, Mendelssohn...",
answer the following question: "What is the
    occupation of Bach?"
Answer: composer; conductor; virtuoso;
    concertmaster; organist; choir director;
violinist; school teacher; harpsichordist;
    musicologist; music educator

###INPUT
Person: "Gavinies"
Sentence: "Gavinies published three books of
     sonatas, and several concertos,
which are very highly esteemed by
    connoisseurs."
Context:
{
 'TheQuarterlyMusicalMagazineAndReview_1825
    -025.txt_p349':
  'It was during this time that he composed,
      as if by inspiration, the celebrated
   Romance de Gavinies, so long in vogue.
      This romance he performed at the age
   of 63, with such exquisite expression as
      to draw tears from a crowded audience
   at a public concert. Gavinies published
      three books of sonatas, and several
   concertos, which are very highly esteemed
       by connoisseurs. A year before his
   death (in 1799)
   he published a work
   entitled "Les
   vingt quatre
   Matine\'es,"
   which contains music of still more
      difficult execution than the Caprices
      of
   Locatelli and Fiorillo. Gavinies was
      particularly celebrated for his skill
      in
   accompaniment, and the taste he displayed
       in the variations he introduced.',
 'TheQuarterlyMusicalMagazineAndReview_1825
    -025.txt_p352':
  'Pagin, born in 1730, went into Italy
     purposely to profit by the instructions
   of Tartini. He returned to Paris in his
      twentieth year, and performed several
   times with great applause at the Concerts
       Spirituel. He however executed
   nothing but the music of his master
      Tartini, which raised a cabal against
      him
   amongst the French musicians, and he one
      day received such ironical applause
   from them, as induced him to quit the
      Concert Spirituel, and he accordingly
   accepted a situation in the house of the
      Count de Clermont.',
 'TheQuarterlyMusicalMagazineAndReview_1825
    -025.txt_p46':
  'Maria Von Weber ... Three Grand Sonatas
     for the Piano Forte, composed by
   Charles Ambrose ...'
}
Question: "What is the occupation of
    Gavinies?"

###ANSWER_PREFIX
```

```
Answer:
```

Listing 1: Full prompt instance with block markers (#)

.2 Tables for Overall Property Extraction

In this appendix, we report the tables with the results of our overall property extraction. We report results for the QID-KG in Table 10, and for the NIL-KG in Table 11.

.3 Tables for Date of Birth Evaluation

In this appendix, we report the tables with the results of our date of birth evaluation. We report results for QID-KG in Table 12, and for NIL-KG in Table 13.

.4 Tables for Occupation Per-sample Evaluation

In this appendix, we report the tables with the results of our occupation evaluation. We report results for QID-KG in Table 14, and for NIL-KG in Table 15.

Table 10. Precision (P), recall (R), and F1 for the six core properties on the **QID-KG**. We evaluate Mixtral-8×7B, Phi-3-Medium, Gemma-2-27B-IT, Llama-3.3-70B, Qwen-2.5-72B, and GPT-4o-mini. Date of birth (DoB) is scored at *year* granularity, while *occupation* is counted correct when at least one predicted occupation QID overlaps with the gold set.

Model	Retr.	CoC			DoB			FamilyName			GivenName			occupation			sexGender		
		F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Boyer-m	Combined	**0.47**	**0.49**	0.45	0.33	0.91	0.20	0.87	0.89	**0.85**	0.44	0.52	0.38	0.18	0.81	0.10	0.99	0.99	0.99
Mixtral	Boyer-M	0.28	0.32	0.25	0.48	0.70	0.36	0.84	0.90	0.80	0.30	0.46	0.22	0.19	0.54	0.11	0.92	0.99	0.86
	-	0.16	0.25	0.12	0.59	0.80	0.47	0.74	0.90	0.63	0.15	0.45	0.09	0.69	0.77	0.63	0.41	0.98	0.26
	BGE	0.17	0.24	0.13	0.43	0.86	0.29	0.72	0.90	0.60	0.21	0.49	0.14	0.73	0.79	0.68	0.69	**1.00**	0.52
	BM25	0.13	0.23	0.09	0.40	0.91	0.26	0.72	0.90	0.60	0.21	0.50	0.13	0.74	0.78	0.69	0.68	0.99	0.52
	Inst	0.15	0.24	0.11	0.40	0.90	0.26	0.73	0.91	0.61	0.24	0.51	0.15	0.71	0.75	0.68	0.68	**1.00**	0.52
	Contr	0.13	0.22	0.09	0.44	**0.92**	0.29	0.73	0.90	0.61	0.19	0.43	0.12	0.75	0.78	0.71	0.68	**1.00**	0.52
	GTR-X	0.15	0.23	0.11	0.42	0.89	0.28	0.75	**0.92**	0.64	0.23	0.54	0.15	0.74	0.77	0.71	0.71	**1.00**	0.55
	GTR-L	0.17	0.25	0.13	0.42	0.91	0.27	0.73	0.90	0.61	0.24	0.55	0.16	0.73	0.77	0.69	0.72	**1.00**	0.57
Phi-3	Boyer-M	0.41	0.44	0.39	0.19	0.37	0.13	0.86	0.89	0.83	0.31	0.41	0.24	0.46	0.77	0.33	0.97	0.97	0.97
	-	0.22	0.28	0.19	0.25	0.55	0.16	0.85	0.89	0.81	0.13	0.45	0.08	0.54	0.77	0.41	0.87	0.96	0.80
	BGE	0.28	0.28	0.27	0.44	0.52	0.39	0.85	0.88	0.82	0.33	0.47	0.26	0.72	0.73	0.70	0.94	0.95	0.93
	BM25	0.29	0.29	0.29	0.45	0.54	0.38	0.83	0.87	0.80	0.31	0.44	0.24	0.71	0.72	0.70	0.94	0.96	0.93
	Inst	0.29	0.29	0.29	0.44	0.54	0.37	0.84	0.87	0.81	0.32	0.45	0.25	0.70	0.72	0.68	0.94	0.95	0.93
	Contr	0.30	0.31	0.30	0.45	0.53	0.39	0.85	0.88	0.82	0.32	0.45	0.25	0.75	0.76	0.73	0.93	0.94	0.92
	GTR-X	0.29	0.29	0.29	0.46	0.56	0.39	0.85	0.88	0.82	0.33	0.46	0.26	0.72	0.74	0.71	0.95	0.95	0.94
	GTR-L	0.29	0.30	0.29	0.45	0.54	0.39	0.85	0.88	0.82	0.32	0.44	0.25	0.72	0.73	0.71	0.94	0.95	0.93
Gemma	Boyer-M	0.40	0.41	0.39	0.31	0.46	0.24	0.87	0.90	0.84	0.36	0.48	0.29	0.44	0.75	0.32	0.99	0.99	0.99
	-	0.28	0.28	0.28	0.54	0.72	0.43	0.86	0.89	0.82	0.25	0.45	0.17	0.74	0.75	0.73	0.97	0.99	0.95
	BGE	0.31	0.31	0.31	0.54	0.67	0.46	0.86	0.89	0.84	0.40	0.55	0.31	0.79	0.79	0.78	0.98	0.99	0.98
	BM25	0.31	0.31	0.31	0.56	0.71	0.47	0.86	0.90	0.83	0.38	0.54	0.29	0.80	0.80	**0.79**	0.98	0.99	0.98
	Inst	0.32	0.32	0.32	0.55	0.69	0.46	0.86	0.89	0.84	0.40	0.55	0.32	0.79	0.80	0.78	0.98	0.99	0.98
	Contr	0.30	0.30	0.30	0.54	0.67	0.45	0.86	0.90	0.83	0.40	0.56	0.32	0.80	0.80	**0.79**	0.98	0.99	0.97
	GTR-X	0.30	0.30	0.30	0.55	0.66	0.46	0.87	0.90	0.84	0.40	0.55	0.31	0.79	0.80	**0.79**	0.98	0.99	0.98
	GTR-L	0.31	0.31	0.31	0.55	0.68	0.46	0.87	0.89	0.84	0.40	0.56	0.32	0.79	0.80	0.78	0.99	0.99	0.98
Llama	Boyer-M	**0.47**	0.48	**0.47**	0.52	0.80	0.39	0.87	0.89	0.84	0.52	0.60	0.46	0.22	0.85	0.13	0.99	0.99	0.99
	-	0.37	0.37	0.37	**0.62**	0.62	**0.61**	0.87	0.90	0.84	0.52	**0.67**	0.42	**0.81**	0.84	**0.79**	0.99	0.99	0.99
	BGE	0.38	0.38	0.38	0.60	0.61	0.60	0.87	0.90	**0.85**	0.58	0.65	0.53	0.77	0.85	0.71	0.99	0.99	0.99
	BM25	0.39	0.39	0.39	0.61	0.62	**0.61**	0.87	0.91	**0.85**	**0.59**	0.65	**0.54**	0.79	0.85	0.73	0.99	0.99	0.99
	Inst	0.38	0.38	0.38	0.60	0.61	0.60	**0.88**	0.91	**0.85**	0.57	0.64	0.52	0.76	0.84	0.70	0.99	0.99	0.99
	Contr	0.38	0.39	0.38	0.61	0.62	0.60	0.87	0.90	**0.85**	0.58	0.64	0.53	**0.81**	**0.86**	0.76	0.99	0.99	0.99
	GTR-X	0.37	0.37	0.37	0.61	0.61	0.60	0.87	0.90	0.84	**0.59**	0.66	**0.54**	0.79	**0.86**	0.73	0.99	0.99	0.99
	GTR-L	0.39	0.39	0.39	0.60	0.61	0.60	0.87	0.90	**0.85**	0.58	0.65	0.52	0.77	0.84	0.72	0.99	0.99	0.99
Qwen	Boyer-M	**0.47**	0.47	**0.47**	0.36	0.76	0.24	0.87	0.90	0.84	0.37	0.52	0.28	0.22	0.56	0.14	0.98	0.99	0.98
	-	0.36	0.36	0.36	0.58	0.61	0.55	0.86	0.90	0.83	0.23	0.54	0.14	0.67	0.72	0.62	0.97	0.99	0.96
	BGE	0.37	0.37	0.37	0.55	0.58	0.53	0.87	0.90	0.84	0.39	0.59	0.29	0.70	0.75	0.66	0.98	0.99	0.98
	BM25	0.36	0.36	0.36	0.58	0.61	0.54	0.87	0.90	0.84	0.39	0.57	0.29	0.68	0.72	0.64	0.99	0.99	0.98
	Inst	0.37	0.37	0.37	0.58	0.62	0.55	0.87	0.90	0.84	0.39	0.55	0.31	0.68	0.73	0.65	0.99	0.99	0.98
	Contr	0.37	0.37	0.37	0.59	0.63	0.56	0.87	0.90	0.84	0.39	0.59	0.30	0.70	0.74	0.67	0.98	0.99	0.98
	GTR-X	0.39	0.39	0.38	0.56	0.57	0.55	0.87	0.90	**0.85**	0.41	0.57	0.32	0.70	0.75	0.65	0.99	0.99	0.99
	GTR-L	0.38	0.38	0.38	0.57	0.61	0.54	0.87	0.90	0.84	0.41	0.58	0.32	0.71	0.75	0.67	0.99	0.99	0.98
Gpt-4o	Boyer-M	0.28	0.29	0.28	0.41	0.84	0.27	0.86	0.88	0.83	0.34	0.45	0.27	0.16	0.47	0.10	0.99	0.99	0.99
	-	0.19	0.19	0.19	0.54	0.55	0.53	0.86	0.89	0.83	0.32	0.52	0.23	0.73	0.75	0.71	0.99	0.99	0.98
	BGE	0.22	0.22	0.22	0.54	0.55	0.53	0.86	0.89	0.84	0.39	0.52	0.31	0.78	0.79	0.77	0.99	0.99	0.99
	BM25	0.23	0.23	0.23	0.56	0.57	0.55	0.86	0.89	0.83	0.38	0.52	0.30	0.80	0.81	**0.79**	0.99	0.99	0.99
	Inst	0.24	0.24	0.24	0.55	0.56	0.54	0.86	0.89	0.83	0.40	0.54	0.31	0.78	0.78	0.77	**1.00**	**1.00**	**1.00**
	Contr	0.22	0.22	0.22	0.55	0.56	0.54	0.86	0.89	0.84	0.42	0.56	0.33	0.78	0.80	0.77	0.99	0.99	0.99
	GTR-X	0.25	0.25	0.25	0.55	0.56	0.54	0.87	0.90	0.84	0.41	0.55	0.33	0.77	0.78	0.76	0.99	0.99	0.99
	GTR-L	0.25	0.25	0.25	0.56	0.57	0.55	0.87	0.90	0.84	0.41	0.56	0.33	0.78	0.79	0.78	0.99	0.99	0.99

Table 11. Precision (P), recall (R), and F1 for the six core properties on the **NIL-KG**. We evaluate Mixtral-8×7B, Phi-3-Medium, Gemma-2-27B-IT, Llama-3.3-70B, Qwen-2.5-72B and GPT-4o-mini. Date of birth (DoB) is evaluated at *year* granularity; an *occupation* prediction is treated as correct when it shares at least one occupation QID with the gold labels.

Model	Retr.	CoC			DoB			FamilyName			GivenName			occupation			sexGender		
		F1	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
Boyer-m	Combined	0.23	0.21	0.25	0.11	**1.00**	0.06	0.85	0.85	0.86	0.37	0.30	**0.47**	0.29	**0.72**	0.18	0.92	0.87	0.97
Mixtral	Boyer-M	0.24	0.32	0.19	0.15	0.20	0.12	0.87	0.91	0.84	0.45	0.47	0.44	0.36	0.64	0.25	0.84	0.91	0.77
	-	0.17	0.29	0.11	0.15	0.22	0.12	0.82	0.93	0.72	0.42	0.56	0.33	0.48	0.59	0.41	0.35	0.92	0.22
	BGE	0.19	0.48	0.11	0.11	**1.00**	0.06	0.82	**0.94**	0.73	0.41	0.50	0.35	0.54	0.58	0.50	0.56	0.89	0.41
	BM25	0.12	0.67	0.07	0.11	**1.00**	0.06	0.82	0.92	0.74	0.41	0.50	0.35	0.53	0.58	0.50	0.60	0.92	0.45
	Inst	0.15	0.47	0.09	0.11	0.50	0.06	0.80	**0.94**	0.69	0.46	0.56	0.39	0.55	0.60	0.51	0.61	0.90	0.47
	Contr	0.10	0.50	0.06	0.00	0.00	0.00	0.82	0.92	0.74	0.45	0.55	0.39	0.56	0.61	0.51	0.66	**0.94**	0.50
	GTR-X	0.19	0.56	0.11	0.00	0.00	0.00	0.79	0.92	0.69	0.43	0.51	0.37	0.52	0.59	0.46	0.60	0.92	0.45
	GTR-L	0.19	0.62	0.11	0.11	**1.00**	0.06	0.82	0.92	0.73	0.46	0.56	0.39	0.47	0.53	0.42	0.64	**0.94**	0.49
Phi-3	Boyer-M	0.27	0.26	0.28	0.03	0.02	0.06	0.85	0.86	0.85	0.39	0.35	0.45	0.41	0.63	0.31	0.92	0.87	0.97
	-	0.29	0.30	0.28	0.07	0.08	0.06	0.88	0.91	0.86	0.27	**0.59**	0.18	0.30	0.46	0.22	0.84	0.86	0.82
	BGE	0.29	0.26	0.32	0.04	0.03	0.06	0.84	0.86	0.83	0.43	0.42	0.44	0.48	0.49	0.47	0.88	0.84	0.92
	BM25	0.30	0.27	0.33	0.08	0.06	0.12	0.86	0.89	0.84	0.43	0.42	0.44	0.44	0.43	0.44	0.91	0.86	0.95
	Inst	0.35	0.32	0.38	0.09	0.07	0.12	0.88	0.89	**0.87**	0.44	0.44	0.44	0.42	0.42	0.42	0.87	0.83	0.90
	Contr	0.29	0.26	0.33	0.04	0.03	0.06	0.88	0.91	0.85	0.42	0.42	0.42	0.45	0.45	0.46	0.90	0.86	0.94
	GTR-X	0.31	0.28	0.36	0.04	0.03	0.06	0.88	0.92	0.85	0.42	0.42	0.42	0.43	0.44	0.43	0.89	0.86	0.92
	GTR-L	0.31	0.28	0.36	0.08	0.06	0.12	0.84	0.86	0.83	0.43	0.42	0.44	0.48	0.48	0.48	0.88	0.86	0.90
Gemma	Boyer-M	0.22	0.21	0.24	0.03	0.02	0.06	0.86	0.87	0.85	0.43	0.44	0.42	0.40	0.62	0.30	0.91	0.86	0.96
	-	0.31	0.28	0.36	0.07	0.08	0.06	**0.90**	0.93	**0.87**	0.47	0.52	0.42	0.51	0.53	0.50	0.83	0.82	0.83
	BGE	0.28	0.25	0.32	0.06	0.07	0.06	0.89	0.92	0.86	0.45	0.49	0.42	0.61	0.61	0.60	0.89	0.86	0.92
	BM25	0.24	0.22	0.28	0.12	0.13	0.12	0.88	0.89	0.86	0.46	0.51	0.42	0.62	0.61	0.63	0.90	0.86	0.95
	Inst	0.28	0.25	0.32	**0.18**	0.19	0.18	0.88	0.91	0.86	0.46	0.50	0.42	0.60	0.58	0.61	0.90	0.87	0.94
	Contr	0.24	0.21	0.28	0.11	0.10	0.12	0.87	0.90	0.85	0.45	0.48	0.42	0.62	0.61	0.63	0.89	0.85	0.92
	GTR-X	0.27	0.24	0.31	0.12	0.13	0.12	0.88	0.89	0.86	0.45	0.49	0.42	0.58	0.57	0.59	**0.93**	0.88	**0.98**
	GTR-L	0.28	0.25	0.32	0.09	0.08	0.12	0.88	0.89	0.86	0.45	0.48	0.42	0.57	0.56	0.58	0.92	0.90	0.94
Llama	Boyer-M	0.21	0.19	0.24	0.17	0.17	0.18	0.87	0.89	0.86	0.42	0.39	0.45	0.26	0.70	0.16	**0.93**	0.88	**0.98**
	-	0.29	0.26	0.33	0.05	0.03	0.18	0.88	0.91	0.85	0.46	0.48	0.44	0.55	0.57	0.52	0.92	0.88	0.97
	BGE	0.21	0.19	0.24	0.05	0.03	0.18	0.88	0.91	0.86	0.44	0.42	**0.47**	**0.68**	0.68	**0.68**	0.92	0.88	0.97
	BM25	0.20	0.18	0.23	0.03	0.02	0.12	0.88	0.90	0.86	0.42	0.40	0.44	0.67	0.68	0.66	**0.93**	0.88	**0.98**
	Inst	0.22	0.20	0.25	0.07	0.04	0.24	0.87	0.91	0.84	0.41	0.39	0.44	0.67	0.66	**0.68**	**0.93**	0.88	**0.98**
	Contr	0.20	0.18	0.23	0.05	0.03	0.18	0.88	0.91	0.85	0.42	0.39	0.46	**0.68**	0.68	**0.68**	**0.93**	0.88	**0.98**
	GTR-X	0.21	0.19	0.24	0.03	0.02	0.12	0.87	0.90	0.85	0.43	0.41	0.46	0.64	0.64	0.65	**0.93**	0.88	**0.98**
	GTR-L	0.21	0.19	0.24	0.07	0.04	0.24	0.87	0.90	0.84	0.41	0.39	0.44	0.67	0.66	**0.68**	**0.93**	0.88	**0.98**
Qwen	Boyer-M	0.21	0.19	0.24	0.05	0.04	0.06	0.86	0.88	0.85	0.48	0.49	**0.47**	0.37	0.57	0.28	0.92	0.88	0.97
	-	0.22	0.20	0.25	0.05	0.03	0.12	0.88	0.91	0.86	0.50	0.57	0.44	0.43	0.47	0.40	0.89	0.87	0.91
	BGE	0.19	0.17	0.22	0.05	0.03	0.12	0.87	0.90	0.85	0.49	0.52	0.46	0.52	0.52	0.52	0.92	0.89	0.95
	BM25	0.23	0.21	0.26	0.05	0.03	0.12	0.86	0.89	0.84	**0.51**	0.55	**0.47**	0.54	0.54	0.54	0.92	0.89	0.96
	Inst	0.22	0.20	0.25	0.07	0.04	0.18	0.86	0.89	0.84	0.48	0.53	0.44	0.56	0.56	0.56	0.92	0.88	0.96
	Contr	0.23	0.21	0.26	0.06	0.04	0.18	0.88	0.90	0.86	0.48	0.53	0.44	0.52	0.52	0.52	0.92	0.88	0.96
	GTR-X	0.21	0.19	0.24	0.11	0.06	**0.29**	0.86	0.89	0.84	0.48	0.52	0.44	0.56	0.55	0.56	0.92	0.88	0.97
	GTR-L	0.22	0.20	0.25	0.07	0.04	0.18	0.88	0.89	0.86	0.45	0.45	0.44	0.52	0.53	0.51	0.91	0.87	0.95
Gpt-4o	Boyer-M	**0.75**	**0.68**	0.84	0.09	0.20	0.06	0.86	0.89	0.84	0.44	0.44	0.44	0.35	0.53	0.26	0.92	0.87	0.97
	-	0.70	0.62	0.80	0.04	0.02	0.12	0.88	0.91	0.86	0.44	0.46	0.42	0.46	0.51	0.43	0.88	0.85	0.90
	BGE	0.61	0.54	0.70	0.03	0.02	0.12	0.89	0.92	0.86	0.43	0.44	0.42	0.51	0.51	0.50	0.92	0.87	0.96
	BM25	0.58	0.52	0.67	0.05	0.03	0.18	0.88	0.91	0.85	0.45	0.45	0.44	0.64	0.64	0.64	0.91	0.87	0.96
	Inst	0.68	0.61	0.78	0.05	0.03	0.18	0.88	0.91	0.85	0.45	0.46	0.44	0.54	0.53	0.54	0.92	0.87	0.96
	Contr	0.74	0.66	**0.85**	0.05	0.03	0.18	0.88	0.91	0.85	0.45	0.45	0.44	0.56	0.55	0.56	0.92	0.88	0.97
	GTR-X	0.61	0.54	0.69	0.05	0.03	0.18	0.87	0.90	0.85	0.43	0.44	0.42	0.50	0.50	0.50	0.92	0.88	0.97
	GTR-L	0.62	0.55	0.71	0.05	0.03	0.18	0.88	0.91	0.85	0.43	0.44	0.42	0.56	0.55	0.56	**0.93**	0.88	**0.98**

Table 12. DoB performance metrics at different granularity levels for entity linking on known entities (QID). We evaluate Mixtral-8×7B, Phi-3-Medium, Gemma-2-27B-IT, Llama-3.3-70B, Qwen-2.5-72B, and GPT-4o-mini.

Model	Retr.	Exact			Year			Decade			Century		
		F1	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
Boyer-M	Boyer-M	0.27	0.75	0.16	0.32	0.91	0.20	0.34	0.96	0.21	0.36	**1.00**	0.22
Mixtral	Boyer-M	0.26	0.39	0.20	0.48	0.70	0.36	0.53	0.78	0.40	0.62	0.92	0.47
	-	0.34	0.47	0.27	0.59	0.81	0.46	0.62	0.85	0.48	0.68	0.94	0.54
	BGE	0.27	0.53	0.18	0.43	0.86	0.29	0.46	0.92	0.31	0.49	0.98	0.33
	BM25	0.26	0.58	0.17	0.40	0.91	0.26	0.42	0.94	0.27	0.44	0.98	0.28
	Inst	0.25	0.57	0.16	0.40	0.90	0.26	0.42	0.93	0.27	0.44	0.99	0.28
	Contr	0.29	0.62	0.19	0.44	0.92	0.29	0.45	0.95	0.29	0.46	0.98	0.30
	GTR-X	0.26	0.54	0.17	0.42	0.89	0.28	0.43	0.91	0.28	0.46	0.97	0.30
	GTR-L	0.27	0.59	0.18	0.42	0.91	0.27	0.43	0.94	0.28	0.46	0.99	0.30
Phi-3	Boyer-M	0.11	0.22	0.08	0.19	0.36	0.13	0.24	0.46	0.16	0.42	0.80	0.29
	-	0.05	0.11	0.03	0.25	0.56	0.16	0.29	0.64	0.18	0.36	0.82	0.23
	BGE	0.20	0.24	0.18	0.44	0.51	0.39	0.51	0.59	0.44	0.71	0.82	0.62
	BM25	0.20	0.24	0.17	0.44	0.53	0.38	0.51	0.61	0.44	0.70	0.84	0.60
	Inst	0.19	0.24	0.16	0.44	0.53	0.37	0.50	0.61	0.43	0.67	0.82	0.57
	Contr	0.20	0.23	0.17	0.45	0.53	0.39	0.52	0.61	0.45	0.71	0.84	0.62
	GTR-X	0.20	0.24	0.17	0.46	0.56	0.39	0.52	0.63	0.44	0.69	0.83	0.58
	GTR-L	0.20	0.24	0.17	0.45	0.54	0.38	0.52	0.62	0.44	0.70	0.84	0.60
Gemma	Boyer-M	0.19	0.28	0.15	0.31	0.45	0.23	0.37	0.55	0.28	0.54	0.80	0.41
	-	0.31	0.41	0.24	0.54	0.72	0.43	0.60	0.80	0.48	0.68	0.92	0.55
	BGE	0.29	0.35	0.24	0.54	0.67	0.46	0.63	0.77	0.53	0.75	0.92	0.63
	BM25	0.30	0.38	0.25	0.56	0.70	0.47	0.62	0.78	0.52	0.75	0.93	0.62
	Inst	0.28	0.35	0.24	0.55	0.68	0.46	0.62	0.76	0.52	0.74	0.92	0.62
	Contr	0.30	0.37	0.25	0.54	0.67	0.45	0.61	0.76	0.51	0.74	0.92	0.62
	GTR-X	0.28	0.34	0.24	0.54	0.66	0.46	0.63	0.76	0.53	0.76	0.92	0.65
	GTR-L	0.29	0.35	0.24	0.54	0.67	0.45	0.63	0.78	0.53	0.75	0.93	0.63
Llama	Boyer-M	**0.38**	0.58	0.28	0.52	0.81	0.39	0.57	0.89	0.42	0.62	0.96	0.46
	-	0.37	0.37	0.36	**0.62**	0.63	**0.61**	**0.71**	0.72	**0.70**	**0.91**	0.92	0.89
	BGE	0.37	0.37	0.36	0.60	0.60	0.59	0.70	0.71	0.69	0.88	0.89	0.88
	BM25	0.37	0.38	**0.37**	0.61	0.62	**0.61**	0.70	0.71	**0.70**	**0.91**	0.92	**0.90**
	Inst	0.36	0.36	0.36	0.60	0.60	0.59	0.68	0.69	0.68	0.89	0.90	0.88
	Contr	0.37	0.37	0.36	0.61	0.62	0.60	0.69	0.70	0.68	0.90	0.91	0.89
	GTR-X	0.37	0.38	**0.37**	0.60	0.61	0.60	0.70	0.71	**0.70**	0.88	0.89	0.88
	GTR-L	0.37	0.37	**0.37**	0.60	0.61	0.59	0.69	0.69	0.68	0.90	0.91	0.89
Qwen	Boyer-M	0.27	0.58	0.18	0.35	0.75	0.23	0.38	0.81	0.25	0.45	0.95	0.29
	-	0.30	0.32	0.28	0.58	0.62	0.55	0.66	0.71	0.62	0.85	0.91	0.80
	BGE	0.28	0.30	0.27	0.55	0.58	0.53	0.67	0.71	0.64	0.88	0.92	0.84
	BM25	0.30	0.31	0.28	0.58	0.61	0.54	0.69	0.73	0.65	0.89	0.94	0.84
	Inst	0.29	0.31	0.27	0.58	0.62	0.55	0.67	0.72	0.63	0.87	0.93	0.82
	Contr	0.30	0.32	0.29	0.59	0.62	0.56	0.69	0.73	0.65	0.88	0.93	0.84
	GTR-X	0.28	0.28	0.27	0.56	0.57	0.55	0.68	0.69	0.67	0.90	0.91	0.88
	GTR-L	0.29	0.31	0.27	0.57	0.61	0.54	0.67	0.71	0.63	0.87	0.93	0.83
GPT-4o	Boyer-M	0.29	0.61	0.19	0.41	0.85	0.27	0.44	0.92	0.29	0.47	0.97	0.31
	-	0.31	0.32	0.30	0.55	0.57	0.53	0.62	0.64	0.60	0.81	0.83	0.78
	BGE	0.29	0.29	0.28	0.54	0.55	0.53	0.65	0.66	0.64	0.87	0.89	0.86
	BM25	0.30	0.31	0.30	0.56	0.57	0.55	0.64	0.65	0.63	0.88	0.90	0.87
	Inst	0.29	0.29	0.28	0.55	0.56	0.54	0.65	0.65	0.64	0.87	0.88	0.86
	Contr	0.30	0.30	0.29	0.55	0.56	0.54	0.66	0.67	0.65	0.88	0.89	0.86
	GTR-X	0.29	0.30	0.29	0.55	0.56	0.54	0.65	0.66	0.64	0.88	0.90	0.87
	GTR-L	0.31	0.31	0.30	0.56	0.57	0.55	0.66	0.67	0.65	0.87	0.88	0.86

Table 13. DoB performance metrics at different granularity levels for NIL identification (entities not in the knowledge base). We evaluate Mixtral-8×7B, Phi-3-Medium, Gemma-2-27B-IT, Llama-3.3-70B, Qwen-2.5-72B, and GPT-4o-mini.

Model	Retr.	Exact			Year			Decade			Century		
		F1	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
Boyer-M	Boyer-M	0.11	**1.00**	0.06	0.11	**1.00**	0.06	0.11	**1.00**	0.06	0.11	**1.00**	0.06
Mixtral	Boyer-M	0.07	0.10	0.06	0.15	0.20	0.12	0.15	0.20	0.12	0.30	0.40	0.24
	-	0.08	0.11	0.06	0.15	0.22	0.12	0.15	0.22	0.12	0.31	0.44	0.24
	BGE	0.11	**1.00**	0.06	0.11	**1.00**	0.06	0.11	**1.00**	0.06	0.11	**1.00**	0.06
	BM25	0.11	**1.00**	0.06	0.11	**1.00**	0.06	0.11	**1.00**	0.06	0.11	**1.00**	0.06
	Inst	0.11	0.50	0.06	0.11	0.50	0.06	0.11	0.50	0.06	0.11	0.50	0.06
	Contr	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	GTR-X	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	GTR-L	0.11	**1.00**	0.06	0.11	**1.00**	0.06	0.11	**1.00**	0.06	0.11	**1.00**	0.06
Phi-3	Boyer-M	0.03	0.02	0.06	0.03	0.02	0.06	0.06	0.04	0.12	0.22	0.15	0.41
	-	0.07	0.08	0.06	0.07	0.08	0.06	0.07	0.08	0.06	0.34	0.42	0.29
	BGE	0.04	0.03	0.06	0.04	0.03	0.06	0.08	0.06	0.12	0.20	0.15	0.29
	BM25	0.00	0.00	0.00	0.08	0.06	0.12	0.16	0.12	0.24	0.28	0.21	0.41
	Inst	0.00	0.00	0.00	0.09	0.07	0.12	0.09	0.07	0.12	0.18	0.15	0.24
	Contr	0.04	0.03	0.06	0.04	0.03	0.06	0.04	0.03	0.06	0.30	0.22	0.47
	GTR-X	0.04	0.03	0.06	0.04	0.03	0.06	0.04	0.03	0.06	0.16	0.12	0.24
	GTR-L	0.04	0.03	0.06	0.08	0.06	0.12	0.12	0.09	0.18	0.19	0.14	0.29
Gemma	Boyer-M	0.03	0.02	0.06	0.03	0.02	0.06	0.03	0.02	0.06	0.12	0.09	0.24
	-	0.07	0.08	0.06	0.07	0.08	0.06	0.07	0.08	0.06	0.13	0.15	0.12
	BGE	0.06	0.07	0.06	0.06	0.07	0.06	0.06	0.07	0.06	0.19	0.20	0.18
	BM25	0.06	0.07	0.06	0.12	0.13	0.12	0.12	0.13	0.12	0.19	0.20	0.18
	Inst	0.18	0.19	0.18	0.18	0.19	0.18	0.18	0.19	0.18	0.30	0.31	0.29
	Contr	0.11	0.10	0.12	0.11	0.10	0.12	0.11	0.10	0.12	0.26	0.24	0.29
	GTR-X	0.06	0.07	0.06	0.12	0.13	0.12	0.19	0.20	0.18	0.19	0.20	0.18
	GTR-L	0.05	0.04	0.06	0.09	0.08	0.12	0.09	0.08	0.12	0.23	0.19	0.29
Llama	Boyer-M	0.11	0.11	0.12	0.17	0.17	0.18	0.29	0.28	0.29	0.29	0.28	0.29
	-	0.03	0.02	0.12	0.05	0.03	0.18	0.12	0.07	0.41	0.19	0.11	0.65
	BGE	0.03	0.02	0.12	0.05	0.03	0.18	0.17	0.10	0.59	0.24	0.14	0.82
	BM25	0.02	0.01	0.06	0.03	0.02	0.12	0.19	0.11	0.65	0.24	0.14	0.82
	Inst	0.03	0.02	0.12	0.07	0.04	0.24	0.14	0.08	0.47	0.24	0.14	0.82
	Contr	0.03	0.02	0.12	0.05	0.03	0.18	0.16	0.09	0.53	0.21	0.12	0.71
	GTR-X	0.02	0.01	0.06	0.03	0.02	0.12	0.09	0.05	0.29	0.24	0.14	0.82
	GTR-L	0.03	0.02	0.12	0.07	0.04	0.24	0.15	0.09	0.53	0.25	0.15	0.88
Qwen	Boyer-M	0.05	0.04	0.06	0.05	0.04	0.06	0.14	0.12	0.18	0.14	0.12	0.18
	-	0.05	0.03	0.12	0.05	0.03	0.12	0.08	0.05	0.18	0.18	0.11	0.41
	BGE	0.05	0.03	0.12	0.05	0.03	0.12	0.17	0.11	0.41	0.27	0.17	0.65
	BM25	0.02	0.02	0.06	0.05	0.03	0.12	0.14	0.09	0.35	0.24	0.15	0.59
	Inst	0.02	0.01	0.06	0.07	0.04	0.18	0.16	0.10	0.41	0.21	0.13	0.53
	Contr	0.04	0.03	0.12	0.06	0.04	0.18	0.17	0.11	0.47	0.22	0.13	0.59
	GTR-X	0.04	0.03	0.12	0.11	0.06	0.29	0.13	0.08	0.35	0.17	0.10	0.47
	GTR-L	0.05	0.03	0.12	0.07	0.04	0.18	0.14	0.09	0.35	0.25	0.16	0.65
GPT-4o	Boyer-M	0.09	0.20	0.06	0.09	0.20	0.06	0.09	0.20	0.06	0.09	0.20	0.06
	-	0.02	0.01	0.06	0.04	0.02	0.12	0.06	0.03	0.18	0.11	0.07	0.35
	BGE	0.02	0.01	0.06	0.03	0.02	0.12	0.03	0.02	0.12	0.24	0.14	0.82
	BM25	0.02	0.01	0.06	0.05	0.03	0.18	0.09	0.05	0.29	0.19	0.11	0.65
	Inst	0.02	0.01	0.06	0.05	0.03	0.18	0.07	0.04	0.24	0.22	0.13	0.76
	Contr	0.02	0.01	0.06	0.05	0.03	0.18	0.09	0.05	0.29	0.26	0.15	0.88
	GTR-X	0.02	0.01	0.06	0.05	0.03	0.18	0.09	0.05	0.29	0.26	0.15	0.88
	GTR-L	0.02	0.01	0.06	0.05	0.03	0.18	0.07	0.04	0.24	0.23	0.13	0.76

Table 14. Occupation performance metrics for entity linking on known entities (QID). We evaluate Mixtral-8×7B, Phi-3-Medium, Gemma-2-27B-IT, Llama-3.3-70B, Qwen-2.5-72B and GPT-4o-mini.

Model	Retr.	F1			Precision			Recall		
		M	**Med**	**SD**	**M**	**Med**	**SD**	**M**	**Med**	**SD**
Majority Voting	Boyer-M	0.06	0.00	0.20	0.11	0.00	0.31	0.05	0.00	0.17
Mixtral	Boyer-M	0.06	0.00	0.20	0.11	0.00	0.31	0.05	0.00	0.17
	-	0.28	0.31	0.26	0.35	0.33	0.33	0.31	0.25	0.33
	BGE	0.30	0.33	0.25	0.37	0.33	0.33	0.33	0.25	0.32
	BM25	0.31	0.33	0.25	0.38	0.33	0.33	0.33	0.25	0.33
	Inst	0.30	0.33	0.25	0.37	0.33	0.33	0.32	0.25	0.33
	Contr	0.32	0.33	0.25	0.40	0.33	0.34	0.34	0.27	0.32
	GTR-X	0.32	0.33	0.25	0.39	0.33	0.33	0.35	0.25	0.33
	GTR-L	0.31	0.33	0.26	0.38	0.33	0.33	0.34	0.27	0.33
Phi-3	Boyer-M	0.19	0.00	0.31	0.33	0.00	**0.47**	0.15	0.00	0.27
	-	0.19	0.00	0.27	0.33	0.00	0.43	0.16	0.00	0.27
	BGE	0.34	0.33	0.29	0.53	**0.50**	0.41	0.31	0.25	0.32
	BM25	0.34	0.33	0.29	**0.54**	**0.50**	0.42	0.31	0.22	0.32
	Inst	0.33	0.33	0.29	0.52	**0.50**	0.42	0.29	0.22	0.32
	Contr	0.35	0.33	0.29	**0.54**	**0.50**	0.41	0.32	0.25	0.32
	GTR-X	0.35	0.33	0.30	0.53	**0.50**	0.42	0.32	0.25	0.33
	GTR-L	0.34	0.33	0.29	0.52	**0.50**	0.41	0.31	0.25	0.32
Gemma	Boyer-M	0.18	0.00	0.30	0.31	0.00	0.46	0.14	0.00	0.26
	-	0.35	0.33	0.28	0.50	**0.50**	0.38	0.34	0.25	0.33
	BGE	0.38	**0.40**	0.26	0.52	**0.50**	0.37	0.37	0.33	0.32
	BM25	**0.39**	**0.40**	0.26	0.52	**0.50**	0.35	0.39	0.33	0.33
	Inst	0.38	**0.40**	0.27	0.51	**0.50**	0.36	0.38	0.33	0.33
	Contr	0.38	**0.40**	0.26	0.50	**0.50**	0.34	0.38	0.33	0.32
	GTR-X	0.38	**0.40**	0.26	0.51	**0.50**	0.36	0.38	0.33	0.33
	GTR-L	0.38	**0.40**	0.26	0.51	**0.50**	0.36	0.38	0.33	0.33
Llama	Boyer-M	0.07	0.00	0.21	0.13	0.00	0.33	0.06	0.00	0.18
	-	0.36	0.36	0.25	0.42	0.33	0.34	0.44	0.40	0.35
	BGE	0.31	0.32	0.25	0.34	0.29	0.32	0.39	0.33	0.36
	BM25	0.32	0.33	0.25	0.36	0.33	0.33	0.41	0.33	0.36
	Inst	0.31	0.33	0.26	0.36	0.33	0.34	0.38	0.33	0.35
	Contr	0.34	0.33	0.24	0.38	0.33	0.32	0.42	0.40	0.35
	GTR-X	0.31	0.33	0.24	0.36	0.33	0.33	0.40	0.33	0.35
	GTR-L	0.31	0.33	0.25	0.36	0.33	0.33	0.40	0.33	0.36
Qwen	Boyer-M	0.08	0.00	0.21	0.13	0.00	0.33	0.06	0.00	0.18
	-	0.28	0.29	0.26	0.33	0.33	0.33	0.32	0.25	0.33
	BGE	0.30	0.33	0.26	0.36	0.33	0.32	0.34	0.25	0.33
	BM25	0.29	0.33	0.26	0.36	0.33	0.33	0.32	0.25	0.33
	Inst	0.29	0.29	0.26	0.36	0.33	0.34	0.32	0.25	0.33
	Contr	0.31	0.33	0.26	0.38	0.40	0.35	0.34	0.29	0.33
	GTR-X	0.30	0.33	0.26	0.35	0.33	0.33	0.33	0.25	0.34
	GTR-L	0.31	0.33	0.26	0.37	0.33	0.34	0.34	0.29	0.34
GPT-4o	Boyer-M	0.06	0.00	0.19	0.10	0.00	0.30	0.05	0.00	0.17
	-	0.33	0.33	0.26	0.39	0.33	0.33	0.36	0.33	0.33
	BGE	0.34	**0.40**	0.23	0.39	0.33	0.29	0.39	0.33	0.33
	BM25	0.37	**0.40**	0.24	0.42	0.33	0.31	0.41	0.33	0.33
	Inst	0.35	**0.40**	0.24	0.40	0.33	0.31	0.40	0.33	0.33
	Contr	0.36	**0.40**	0.25	0.41	0.33	0.31	0.40	0.33	0.32
	GTR-X	0.34	**0.40**	0.24	0.39	0.33	0.30	0.39	0.33	0.33
	GTR-L	0.35	**0.40**	0.24	0.40	0.33	0.30	0.40	0.33	0.33

Table 15. Occupation performance metrics for NIL identification (entities not in the knowledge base). We evaluate Mixtral-8×7B, Phi-3-Medium, Gemma-2-27B-IT, Llama-3.3-70B, Qwen-2.5-72B and GPT-4o-mini.

Model	Retr.	F1			Precision			Recall		
		M	**Med**	**SD**	**M**	**Med**	**SD**	**M**	**Med**	**SD**
Majority Voting	Boyer-M	0.13	0.00	0.30	0.17	0.00	0.37	0.11	0.00	0.28
Mixtral	Boyer-M	0.16	0.00	0.31	0.22	0.00	0.41	0.14	0.00	0.29
	-	0.24	0.00	0.33	0.25	0.00	0.36	0.26	0.00	0.38
	BGE	0.28	0.00	0.33	0.31	0.00	0.39	0.31	0.00	0.39
	BM25	0.28	0.00	0.33	0.31	0.00	0.39	0.32	0.00	0.40
	Inst	0.29	0.00	0.34	0.31	0.00	0.37	0.34	0.00	0.41
	Contr	0.28	0.00	0.32	0.32	0.00	0.39	0.31	0.00	0.39
	GTR-X	0.23	0.00	0.30	0.26	0.00	0.34	0.26	0.00	0.37
	GTR-L	0.23	0.00	0.32	0.26	0.00	0.38	0.26	0.00	0.38
Phi-3	Boyer-M	0.21	0.00	0.35	0.29	0.00	0.45	0.18	0.00	0.33
	-	0.15	0.00	0.31	0.16	0.00	0.34	0.15	0.00	0.34
	BGE	0.27	0.00	0.34	0.33	0.00	0.41	0.27	0.00	0.36
	BM25	0.26	0.00	0.34	0.33	0.00	0.43	0.26	0.00	0.37
	Inst	0.24	0.00	0.33	0.29	0.00	0.40	0.24	0.00	0.36
	Contr	0.26	0.00	0.33	0.33	0.00	0.41	0.27	0.00	0.37
	GTR-X	0.25	0.00	0.33	0.29	0.00	0.40	0.25	0.00	0.35
	GTR-L	0.28	0.00	0.35	0.35	0.00	0.43	0.29	0.00	0.38
Gemma	Boyer-M	0.21	0.00	**0.36**	0.27	0.00	0.44	0.18	0.00	0.34
	-	0.30	0.00	**0.36**	0.34	0.00	0.42	0.31	0.00	0.39
	BGE	0.36	**0.40**	**0.36**	0.39	0.33	0.41	0.39	0.33	0.41
	BM25	0.36	**0.40**	0.34	0.41	0.42	0.40	0.41	0.33	0.41
	Inst	0.35	**0.40**	0.34	0.39	0.33	0.39	0.39	0.33	0.41
	Contr	0.37	**0.40**	**0.36**	0.40	**0.50**	0.39	0.40	0.33	0.41
	GTR-X	0.34	0.33	0.34	0.36	0.33	0.38	0.37	0.29	0.40
	GTR-L	0.33	0.31	0.35	0.36	0.33	0.38	0.37	0.25	0.41
Llama	Boyer-M	0.11	0.00	0.29	0.15	0.00	0.36	0.10	0.00	0.27
	-	0.28	0.00	0.34	0.27	0.00	0.35	0.36	0.00	0.42
	BGE	0.34	0.33	0.32	0.32	0.20	0.35	0.49	**0.50**	0.43
	BM25	0.32	0.33	0.31	0.30	0.20	0.34	0.46	0.33	0.43
	Inst	0.35	0.33	0.34	0.35	0.23	0.37	0.48	0.42	0.43
	Contr	0.34	0.33	0.31	0.30	0.25	0.32	0.49	**0.50**	**0.44**
	GTR-X	0.33	0.31	0.32	0.29	0.20	0.32	0.48	**0.50**	**0.44**
	GTR-L	0.35	0.33	0.33	0.33	0.25	0.35	**0.50**	**0.50**	**0.44**
Qwen	Boyer-M	0.19	0.00	0.33	0.26	0.00	0.44	0.16	0.00	0.30
	-	0.24	0.00	0.34	0.28	0.00	0.40	0.25	0.00	0.37
	BGE	0.30	0.00	0.33	0.35	0.00	0.40	0.32	0.00	0.39
	BM25	0.30	0.25	0.33	0.33	0.23	0.38	0.33	0.18	0.40
	Inst	0.32	0.29	0.35	0.37	0.29	0.40	0.35	0.20	0.40
	Contr	0.30	0.00	0.35	0.34	0.00	0.40	0.33	0.00	0.41
	GTR-X	0.30	0.27	0.32	0.31	0.25	0.35	0.35	0.20	0.40
	GTR-L	0.30	0.00	0.34	0.33	0.00	0.39	0.34	0.00	0.41
GPT-4o	Boyer-M	0.17	0.00	0.32	0.23	0.00	0.42	0.15	0.00	0.31
	-	0.24	0.00	0.33	0.27	0.00	0.38	0.27	0.00	0.38
	BGE	0.28	0.00	0.32	0.32	0.00	0.39	0.31	0.00	0.39
	BM25	0.38	**0.40**	**0.36**	0.42	0.33	0.41	0.41	0.33	0.41
	Inst	0.31	0.24	0.34	0.36	0.33	0.41	0.33	0.18	0.40
	Contr	0.31	0.29	0.33	0.34	0.33	0.38	0.35	0.20	0.41
	GTR-X	0.27	0.00	0.32	0.30	0.00	0.37	0.32	0.00	0.40
	GTR-L	0.30	0.29	0.33	0.34	0.25	0.38	0.37	0.20	0.42