# Generative Information Extraction from Biographical Sources

**Robin Winkle**
University of Potsdam
robin_winkle@outlook.com

**Jörn Kreutel**
Berliner Hochschule für Technik
jkreutel@bht-berlin.de

**Manfred Stede**
University of Potsdam
stede@uni-potsdam.de

## Abstract

Biographical sources, such as literature encyclopedias, encode knowledge about historical figures in textual form. In this paper, we address the task of consolidating structured biographical information about authors from the former German Democratic Republic into a unified database. To this end, we present a generalizable Information Extraction (IE) system based on LLM prompting. Specifically, we compare two midsized open-source models, Qwen-2.5-32B and Llama-3-70B-Instruct, investigate a range of Prompt Engineering (PE) strategies, and propose a semantic similarity-based evaluation metric for open-ended IE. Our experiments on an unpublished annotated subset of biographical texts deliver moderate precision and variable recall, highlighting both the potential and current limitations of generative IE in the Digital Humanities.

## 1 Introduction

Information Extraction (IE) plays an instrumental role in the Digital Humanities and related fields, where the systematic transformation of unstructured or semi-structured texts into machine-readable data is a prerequisite for large-scale cultural and historical analysis.

Our paper is situated in the context of the research project *Forschungsplattform Literarisches Feld DDR* (FLFDDR) which aims to create a bio-bibliographical database that covers approximately 3,400 authors of literary works which have been identified as forming the literary field (Bourdieu, 1992) of the former German Democratic Republic (GDR). By compiling all available published and archival sources regarding those authors –explicitly highlighting conflicting information derived from different sources– the project seeks to enable qualitative and quantitative research on a complete literary field.[1] As of today, data for a subset of those authors, who participated in a study program on literary writing and who sum up to about 10% of the complete corpus,[2] have been successfully collected, using a customized user interface for manual data entry. To extend the database, a semi-automated process will be required, which motivates the focus of this paper on the extraction of biographical information as an instance of the broader IE task.

Recent advances in Natural Language Processing (NLP) have enabled increasingly efficient IE methods, reducing the need for task- or domain-specific parsing rules (Plum et al., 2019). In particular, transformer-based Large Language Models (LLMs) have been used for IE in a generative, prompt-based setting, which we term *Generative IE* (GenIE). While these models are typically optimized for dialogue rather than structured data production (Liu et al., 2024), they have shown strong generalization capabilities across diverse NLP tasks, which has driven their adoption in IE.

In this work, we present a generalizable end-to-end GenIE framework, designed to obtain structured information from a set of natural language texts. Figure 1 provides a high-level representation of the extraction workflow. The input documents (on the left) are supplied to an LLM alongside customizable prompts that specify the format of the structured output (bottom right) and guide the extraction process[3]. The framework supports systematic assessment of Prompt Engineering (PE) strategies and retains source text fragments for all extractions to enable retroactive verifiability and human-in-the-loop validation.

[1]The list of all authors and comments on their compilation are available here: https://ddr-literatur.de/daily-research/korpus-autorinnen-in-der-ddr/

[2]See: *Forschungsplattform Literarisches Feld DDR: Autor*innen, Werke, Netzwerke. Pilotprojekt: Die Student*innen des Instituts für Literatur "Johannes R. Becher" Leipzig*, funded by Deutsche Forschungsgemeinschaft (DFG), project number 419244741, 2019-2023.

[3]The top right represents an alternative, sequential prompting approach explained in Section 5.



*Proceedings of SIGHUM 2026*, pages 311–322

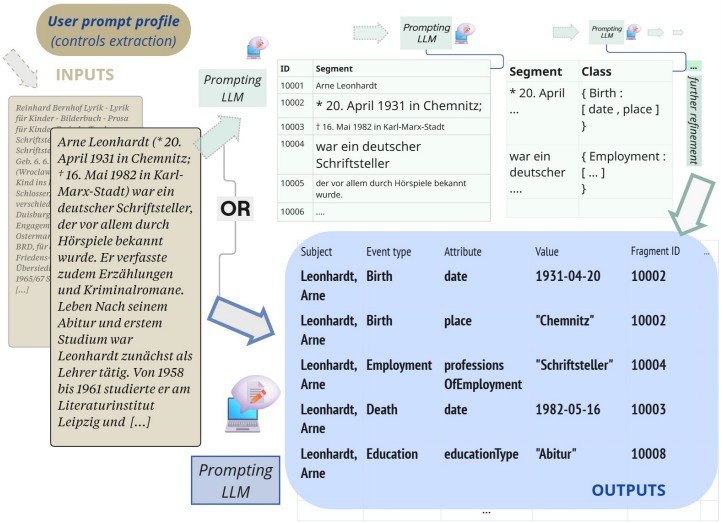


Figure 1: High-level representation of the extraction workflow, including the multi-step (top) vs. single-step approach (bottom).

We assess the framework on an annotated subset of the biographical material curated within the FLFDDR project. Using two moderately sized open-source LLMs, we analyze the impact of different prompting strategies and discuss overall trends and challenges. As an additional contribution, we propose and apply a semantic similarity–based evaluation approach tailored specifically to open IE settings, addressing the limitations of exact-match metrics and restricted GenIE. Our results highlight both the potential and the current limitations of GenIE systems, contributing empirical evidence to ongoing discussions on the responsible and effective use of LLMs in Digital Humanities.

## 2 Background and Related Work

IE is the task of obtaining structured information from unstructured or semi-structured texts written in natural language (McCallum, 2005). During recent decades, computational approaches to IE have enabled the creation of analytically accessible databases for diverse data types, including biographical data (Plum et al., 2019). Within the literary domain, encyclopedias and handbooks, such as those contained in our data set, are among the primary sources for biographical data, listing and contextualizing important life events of literary authors, including a person's place of birth, membership in an institution or the publication of specific works. Structured representations of this data may draw on event-centric data models, which link multiple temporally and contextually connected biographical facts to a unique event within a person's life narrative (Tuominen et al., 2017).[4] Extracting such structured representations from text poses challenges for traditional pipeline-based IE systems, which depend on explicitly modeling intermediate subtasks such as Text Segmentation or Named Entity Recognition (NER).

We address this challenge through a generalizable framework for large-scale IE using pre-trained Large Language Models (LLMs) and Prompt Engineering (PE). While related work has sometimes referred to this approach as "LLM-driven IE", we adopt the term "Generative Information Extraction (GenIE)", originally introduced by Josifoski et al. (2022), since it better reflects the use of generative sequence modeling for structured output. GenIE inherently integrates multiple IE subtasks into an end-to-end sequence generation task (Shen et al., 2023), bypassing the need to explicitly obtain intermediate extraction results. However, it also faces some well-documented challenges associated with generative transformer models in general, such as hallucinations (Huang et al., 2025), biases (Wei et al., 2025), and output variability (Liu et al., 2024).

Related work on GenIE systems has focused particularly on medicine (Goel et al., 2023; Khan et al., 2025; Hu et al., 2024). Gu et al. (2025) evaluate the "out-of-the-box" capabilities of chat-optimised

[4]The concrete event types and attributes employed in the FLFDDR project are motivated by the project's rather wide scope aiming at providing a generic foundation for detailed analyses linking life and work of authors. See table 2 for a full list of event types and attributes from the FLFDDR event data model which have been taken into consideration for the pilot study presented here. For more information on the project's data model, see (Kreutel et al., 2023).

Rodrian, Fred, Chefredakteur, Dir. d. Kinderbuchverl. Berlin; Pr. d. Minist, f.Kultur d. DDR 58 u. 59, Alex-Wedding-Med. 72 u. 76, Johannes-R.-Becher-Med. in Gold 75, Wilhelm- Bracke-Med. 78, Nationalpr. d. DDR (Koll.) 79, Vaterland. Verd.orden in Silber 84, u.a.; * Berlin 14.7.1926, † Berlin 25.5.1985; Kinderbuch u. -film. V: Das Wolkenschaf 58; Das Entenliescl 60; Der Märchenschimmel 60; **[...]** Treffpunkt Erfurt 62; Hirsch Heinrich 65. P: Das Wolkenschaf 64. Lit: DLL, Bd XIII 91; Biogr. Hdb. SBZ/DDR 96, 97.

Figure 2: Source text about author Fred Rodrian from *Kürschners Deutscher Literatur-Kalender Nekrolog 1971-1998* lexicon.

Christine Ernst (* 7. Februar 1938 in Reichenberg, Tschechoslowakei) ist eine deutsche Politikerin (SPD) und ehemaliges Mitglied des Sächsischen Landtages. Christine Ernst besuchte die Grund- und Oberschule in Magdeburg und anschließend die Fachschule für Bibliothekare in Leipzig. Danach war sie als Bibliothekarin und Leiterin eines Jugend-Literatur-Clubs in Magdeburg. Von 1976 bis 1979 folgte ein Studium am Literaturinstitut „Johannes R. Becher“ in Leipzig. Zwischen 1979 und 1980 war sie **[...]**

Figure 3: Source text about author Christine Ernst from *Wikipedia*

models by prompting them to extract medical determinants from text-based health records. Their hypothesis –that general-purpose LLMs can be adopted for IE without task-specific fine-tuning– aligns with the methodology presented in this work. The results of their study suggest that even moderately-sized open-source models substantially outperform naive pattern-matching and can deliver strong performance in IE-related tasks in various domains, particularly when effective PE is applied.

Recent studies have also reported encouraging results with GenIE systems in other domains, including law (Li and Yi, 2024), finance (Li et al., 2025; Kong et al., 2024), and scientific text (Dagdelen et al., 2024). Polak and Morgan (2023) achieve up to 90.8% precision and 87.7% recall in a study on the use of *ChatGPT* for data extraction in a materials science context. However, their formulation evaluates the IE task under a highly constrained matching scheme: extracted outputs are compared against ground truth labels using strict domain-specific equivalence criteria. In contrast, our work targets fully open-ended generative extraction without predefined value sets, and is designed to operate with open-source models and without the need for strict equivalence definitions.

The implementation of our framework is in part informed by these studies. However, to the best of our knowledge, there have not been any systematic attempts to extract structured data from biographical text sources using GenIE systems.

## 3 Data

This study draws on biographical texts compiled in the interdisciplinary FLFDDRproject. The complete corpus contains documents from about 600 different German language sources, including journals, newspapers, online materials, interviews, and around 2,300 articles from 150 literature encyclopedias and handbooks (Kreutel et al., 2023).

Some instances in the source data were digitized using OCR, which may have introduced a slight degree of character-level noise. Other forms of noise may arise from a small amount of inconsistencies between source texts and evaluation data, for example, in the case of author Christine Ernst who was known as Christine Lindner before changing her name. As illustrated by Figures 2-3, the input documents feature a considerable amount of domain-specific abbreviations (e.g. "Lit.büros" for *Literaturbüros*[5]) as well as metadata strings without direct informational value (e.g. "GND: 139456139"). Excluding extreme outliers, the texts contain approximately 1,170 characters on average, with substantial variation between sources.

In our experiments, we use a subset of 188 articles from 72 sources manually annotated by five trained student research assistants under the supervision of literary scholars. Table 5 in the appendix shows the most common text sources in the source corpus and the evaluation subset. The subset primarily features literature encyclopedia[6] articles, a curated and quality-controlled text type containing highly condensed biographical information. As a result, the annotated documents are mostly compact semi-structured entries using a telegraphic style, such as in the example shown in Figure 2. This style is characterized by minimal phrasing with infrequent use of function words and connectors. More narrative texts, such as the one represented by Figure 3, are present but less frequent in the evaluation set.

We process the annotations to fit the format described in Section 4.1. After cleaning, deduplication, and consistency checks, the final evaluation set contains documents labelled with 6,787 information units. The most prevalent event types extracted by annotators capture information relating to authors' education and professional trajectories, besides singular events like birth and death. Attributes such as *begin*, *end*, and *place* that occur

[5]Translation: "Literary Association"
[6]German: *Literaturlexika*

across multiple event types are among the most frequent. Since the annotations emerged out of the project's data collection workflow, they exhibit occasional gaps in coverage and mismatches between surface forms and normalized attribute values, a point we return to in Section 6.2.

## 4 Methods

This section describes the methodological design of our experiments, including the selected output representation, language models, prompting strategies, and a customized evaluation setup.

### 4.1 Output Format

Data models based on the Resource Description Framework (RDF) for biographical data like the one proposed by Tuominen et al. (2017) commonly use knowledge triples that rely on abstract references such as Uniform Resource Identifiers (URI) and entity placeholders to form a graph-like representation. This can present challenges from an LLM generation perspective, since these models are trained for fluent text generation rather than for consistent handling of abstract identifiers. For example, an author may be represented by a URI such as `ex:author_123`, which must be reused consistently across events, a requirement that generative models may fail to satisfy reliably.

| **Field** $f$ | **Field value** $x_f$ |
|---|---|
| SUBJECT | Rodrian, Fred |
| EVENT TYPE | birth |
| ATTRIBUTE | place |
| VALUE | Berlin |
| SOURCE FRAGMENT | geb. am 14.7.1926 im Berliner Osten |

Table 1: Structured information unit associated with text fragment relating to the birth of Fred Rodrian.

Therefore, we adopt a flattened JSON output format with natural language values that is better aligned with the generative strengths of LLMs. A tabular representation of this format is given in Table 1. This format allows for an arbitrary number of tuples, each consisting of a field $f$ and the appropriate field value $x_f$. In our experiments, each extracted information unit is configured to contain the following fields: (1) a SUBJECT (the person concerned), (2) an EVENT TYPE (e.g., *birth*, *education*), (3) an ATTRIBUTE specifying the event further (e.g., *place, institution*), (4) the extracted VALUE encoded as a string, number, or date, and finally, (5) a SOURCE FRAGMENT that grounds the information unit in the original text. Here, the term ATTRIBUTE denotes a property of an EVENT TYPE in the underlying event data model, e.g., the "place" attribute of a "birth" event. While the VALUE is unconstrained, EVENT TYPES and ATTRIBUTES are selected from predefined lists (see Table 2) in our experiments to constrain the extraction scope and improve output consistency.

### 4.2 LLM Selection

To avoid reproducibility issues and access limitations of commercial LLMs, only open-source models were considered. The *Hugging Face* Open-Source Language Model Leaderboard[7], a well-established source with extensive model coverage and diverse benchmark evaluations, serves as the starting point for a multi-stage selection process balancing performance, openness, computational cost, and practical feasibility. Models were ranked according to three evaluation suites relevant to structured information extraction: IFEval (format adherence), MuSR (language understanding), and MMLU-Pro (domain knowledge and reasoning). Models are further evaluated by affordability and environmental footprint. To complement the leaderboard-based shortlist, we examined reports from related IE work to highlight models proven effective in structured extraction.

From this process, two models offering a strong balance of quality and efficiency were selected: *Qwen-2.5-32B*[8] (Yang et al., 2024) and *Llama-3-70B-Instruct*[9] (Dubey et al., 2024). A brief heuristic check on a small sample of biographical texts confirmed that both models produce stable, well-formatted outputs suitable for the GenIE task at hand.

### 4.3 Prompt Engineering

Given its central role in guiding the model's reasoning and extraction decisions, prompt design is handled through a modular setup that automatically assembles prompts by integrating different functional components such as concise instructions, brief descriptions of the input context, examples illustrating the output format, or extraction scope

[7] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, archived, last accessed in June 2025

[8] https://huggingface.co/Qwen/Qwen2.5-32B

[9] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

| Event Types | Attributes |
|---|---|
| Acquaintance | acquaintanceType \| begin \| durationOfEvent \| end \| participants |
| Birth | place \| begin |
| Citizenship | begin \| end \| institution |
| ConfessionAffiliation | begin \| confession \| institution |
| Death | begin \| deathType \| participants \| place |
| Education | begin \| educationType \| educationalContentsOfEducation \| end \| institution \| place |
| Employment | begin \| end \| institution \| place \| professionsOfEmployment |
| Flat | begin \| durationOfEvent \| end \| migratedFromPlace \| place |
| ForcedStay | begin \| durationOfEvent \| end \| forcedStayType \| institution \| place \| politically-Motivated |
| Funding | begin \| durationOfEvent \| end \| fundingInstitution \| place |
| Granting | awardTypeOfGranting \| begin |
| Journey | begin \| destinations \| durationOfEvent \| end \| participants \| place |
| Membership | begin \| durationOfEvent \| end \| institution \| participants \| place |
| MilitaryService | begin \| durationOfEvent \| end \| institution \| participatedWarOfService \| place \| professionsOfEmployment |
| NamedEvent | begin \| durationOfEvent \| end |
| NoEmployment | begin \| durationOfEvent \| end |
| Origin | socialClasses |
| Orphancy | begin |
| Parenthood | begin \| childGender \| childName |
| Surveillance | begin \| durationOfEvent \| end \| place \| surveillingInstitution |

Table 2: Permissible values for extraction fields "Event type" and "Attribute".

constraints. A condensed version of the specific prompt used in our experiments is provided in Table 6. By switching components in or out, the system can be adapted to test various extraction strategies while keeping prompts stable and manageable.

The prompting module also incorporates a set of advanced Prompt Engineering (PE) strategies designed to influence how the model reasons through the extraction task. While general PE aims to optimize prompts to be clear, task-specific and concise, advanced PE targets the model's internal reasoning behavior or attempts to refine the answer space by providing specially crafted output examples.

In this study, three such strategies are implemented. (i) Few-shot prompting (Agarwal et al., 2024) is used to guide the model by analogy, providing concise examples that demonstrate the expected output format and reduce ambiguity in tasks requiring consistent formatting. (ii) Chain-of-Thought (CoT) prompting (Wei et al., 2022) is included as an optional component to encourage explicit intermediate reasoning. And finally, (iii) self-refinement (Madaan et al., 2023) is integrated as a lightweight revision step, allowing the model to critique and correct its own output to improve completeness and adherence to format requirements.

These techniques represent widely adopted model-agnostic methods that can be integrated without external retrieval systems or task-specific fine-tuning. We systematically assess the contribution of these methods to extraction quality through empirical experiments, cf. Section 5.

### 4.4 Evaluation Metric

For open IE tasks, traditional accuracy is ill-defined since the space of true negatives (TN) is difficult to define or possibly unbounded: if a text contains only sparse information about a subject, then an IE system would have to generate a large number of TN (e.g. date of birth: "Unknown", date of death: "Unknown", and so forth) to receive a high evaluation score. Furthermore, strict surface-form matching can be brittle since valid extraction values are not drawn from a closed set in open IE (e.g. if a text cites a person's main occupation as "author", "writer" should also count as a valid extraction value). This contrasts with closed IE, where the task is framed as a multi-choice classification problem (Gu et al., 2025, p. 4) or the TN space is well-bounded at fragment level (Polak and Morgan, 2023, p. 9).

To robustly compare extractions against human annotations, we adopt a type-aware semantic similarity function $\sigma : X \times X \to [0, 1]$ that operates at the level of an extraction field $f$ and determines the similarity between this field's value $x_f$ (cf. Section 4.1) and its counterpart in the annotation set. If $x_f$ is a string literal, cosine similarity is computed using suitable embedding-based representations. This is complemented by specialized similarity scoring functions for cases where $x_f$ is a date or numerical

value.

To estimate the overall alignment score $\phi(e, g)$ between a given candidate unit $e \in \mathcal{E}$ and a gold unit $g \in \mathcal{G}$, let $F$ denote the set of extraction fields required to meet the minimum similarity threshold $\theta_f$ for $f \in F$, and let $w_f$ denote the assigned weight of $f$. Let $x_{f,e}$ and $x_{f,g}$ be the values of $f$ in $e$ and $g$, respectively. The alignment score is then given by Equation 1 if for all $f \in F$: $\sigma_f(x_{f,e}, x_{f,g}) \geq \theta_f$, otherwise it is equal to zero.

$$\phi(e, g) = \frac{\sum_{f \in F} w_f \cdot \sigma_f(x_{f,e}, x_{f,g})}{\sum_{f \in F} w_f} \quad (1)$$

We use $\phi$ to obtain the best precision-oriented alignment $\mu_P(e) = \max_{g \in \mathcal{G}} \phi(e, g)$ and compute soft precision for a set of extracted information units $\mathcal{E}$ as:

$$\text{Precision*} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mu_P(e) \quad (2)$$

We then reverse the comparison to obtain the best recall-oriented alignment $\mu_R(g) = \max_{e \in \mathcal{E}} \phi(e, g)$ and then calculate soft recall for a set of gold information units $\mathcal{G}$ as:

$$\text{Recall*} = \min\left(\frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mu_R(g),\ \frac{|\mathcal{E}|}{|\mathcal{G}|}\right) \quad (3)$$

The upper bound ensures that recall never exceeds the maximum achievable standard recall score, which is equal to the number of candidate elements divided by the number of gold elements, assuming perfect precision.

To summarize, the metric compares the set of extracted units $\mathcal{E}$ against the set of human-annotated units $\mathcal{G}$ by finding, for each element in one set, the closest semantic match in the other set. High scores indicate strong alignment, while low scores signal spurious or unrecalled units. These scores are then aggregated across the entire set to compute overall precision, recall, and F1. Unlike prior work which relies on closed label spaces and employs strict matching, this metric is robust to surface form variability and allows GenIE systems to generate any extraction value that is semantically equivalent. Our evaluation approach is designed to be generalizable and can be adapted to different data sets and domains, since for each field $f$ contained in an extraction unit for a given domain, the similarity function $\sigma_f$, the minimum similarity threshold $\theta_f$ and the weight $w_f$ are configurable.

| Configuration | Model | Heuristically optimized Prompt | Subtask Handling | Reasoning | Number of examples |
|---|---|---|---|---|---|
| BASE | Qwen-2.5 (32B) | Yes | simultaneous | default | 2 |
| EXMP+ | | | | | 13 |
| LARGER-M | Llama-3 (70B) | | | | |
| LARGER-M-EXMP+ | Llama-3 (70B) | | | | 13 |
| MULTI | | | sequential | | |
| CoT | | | | CoT | 13 |

Table 3: Experimental configurations with deviations from the default setup highlighted. Blank entries inherit the default.

## 5 Experiments

We evaluate the proposed framework by comparing several experimental variants to a BASE configuration. These variants, summarized in Table 3, systematically isolate the effect of individual factors, such as model size, number of few-shot examples, multi-step extraction and reasoning strategies.

The BASE configuration employs the standard *Qwen-2.5* model with a single-step prompting setup, default reasoning, and two in-context examples. LARGER-M replaces the baseline model with the larger LLM (cf. Section 4.2) while keeping all other settings fixed. LARGER-M EXMP+ extends this setting by increasing the number of few-shot examples. Analogously, EXMP+ augments BASE by including more output examples without changing model size.

Additionally, MULTI implements a multi-step IE pipeline (see Figure 1, top row), in which the model is sequentially prompted to (i) segment input texts, (ii) extract structured information from segments, and then (iii) perform a single self-refinement pass aimed at improving completeness and strict format adherence. The CoT variant activates explicit CoT instructions, encouraging step-by-step reasoning in the model response.

The performance of the system in the different configurations is assessed using the semantic similarity-based open extraction metric described in Section 4.4, computed against the manually annotated reference data mentioned in Section 3. For our experiments, we set $\theta = 1$ for all extraction fields other than VALUE, effectively enforcing exact matching for these fields since their extraction is constrained by predefined lists (cf. Section 4.1). For string comparison we use an embedding model fine-tuned for cosine similarity computation in German and English[10].

[10] https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer.

## 6 Results

We report the empirical results of our experiments and analyse the behavior of the proposed GenIE framework for the configurations described in Section 5. First, we present a quantitative assessment using the evaluation metric described in Section 4.4 complemented by several supplementary measures, then we conduct a qualitative error analysis that contextualizes observed performance patterns.

### 6.1 Performance

Table 4 summarizes performance at the document level across all configurations. Overall precision was relatively stable (0.61–0.71), while recall varied substantially (0.33–0.75), resulting in F1 scores between 0.45 and 0.68. The strongest configuration, LARGER-M EXMP+, achieved the highest F1 (0.68) driven by a large recall gain (0.75), at the cost of reduced precision (0.62).

Extraction density, i.e. the number of units generated per 100 input characters, shows a clear positive correlation with recall and F1, implying that configurations producing more extraction units consistently achieve higher coverage, despite introducing additional false positives (FP). LARGER-M EXMP+ exhibits the highest extraction density (1.80 units per 100 characters) paired with one of the lowest false negative (FN) rates (6.53%), supporting the intuition that under-generation is more detrimental than moderate over-generation.

Our experiments show that including more examples in the prompt (i.e., few-shot prompting) produced the most substantial performance gains, particularly for the larger model (+0.33 recall from LARGER-M to LARGER-M EXMP+). In contrast, CoT prompting yields only marginal F1 improvements, primarily by trading precision for recall, and multi-step execution with self-refinement degrades performance due to severely reduced extraction density. The direct model comparison reveals that simply using the larger model without prompt optimisation did not improve performance.

Beyond core metrics, the distribution of extraction fields in the system's output largely aligned with the evaluation set, displaying an overall overlap of 60-64%. In both sets, *Education*, *Birth*, and *Employment* events were among the most prevalent event types. However, *Granting* events (i.e., when an author was awarded a prize) were extracted less frequently by the GenIE framework, while *ForcedStay* and *MilitaryService* type events gained relative importance. We note that the system output exhibits a systematic bias towards the extraction of date values (e.g., `13-01-1970`). The percentage of extracted information units containing these values was consistently higher, ranging from 28.3% to 40.2% across configurations, compared to only 9.9% in the gold annotations.

Overall, the results highlight extraction density and example-based prompting as the main drivers of performance, while CoT prompting, multi-step task decomposition, and model scaling in isolation provided limited benefit.

### 6.2 Error Analysis and Discussion

Manual inspection of system outputs[11] reveals that human annotations were frequently incomplete. As a result, many factually correct system extractions absent from the gold data, were classified as FPs. Precision scores therefore need to be interpreted with caution, as they partly reflect annotation coverage rather than true extraction quality. For example, in one document, the system correctly extracted information on the social background of author Adolf Görtz (*son of a factory worker*), but no corresponding annotation exists in the evaluation set. Such cases are penalized as FPs despite being grounded in the text, highlighting an inherent limitation of annotation-based evaluation for open-ended IE.

Beside these "false FPs", a common pattern concerned surface-level variation between extracted values and gold annotations. Here, the evaluation metric proved rather robust since it was specifically designed to account for this type of variation (cf. Section 4.4). For example, in the case of one article that mentioned an author's employment in "Children's and young adult literature", the GenIE system extracted this value literally while human annotators simply noted "Children's books" in the equivalent field. In these cases, the embedding-based metric assigns partial credit based on semantic similarity, yielding high but sub-perfect scores. While this systematically pulls evaluation scores away from full precision, it reflects a fairer assessment than exact-match evaluation, which would fail to credit semantic similarity entirely.

Less frequently, errors arose from metric failures. For instance, the historically equivalent place names (e.g., *Chemnitz* vs. *Karl-Marx-Stadt*) were treated as dissimilar in one case. This type of error

[11] An in-depth example analysis can be viewed on github.com

| Metric/Experiment | MULTI | COT | EXMP+ | LARGER-M | LARGER-M-EXMP+ | BASE |
|---|---|---|---|---|---|---|
| F1 | 0.45 | 0.61 | 0.63 | 0.52 | 0.68 | 0.55 |
| Precision | 0.71 | 0.61 | 0.68 | 0.68 | 0.62 | 0.71 |
| Recall | 0.33 | 0.62 | 0.59 | 0.43 | 0.75 | 0.45 |
| % False positives | 12.43 | 20.03 | 12.89 | 15.70 | 17.59 | 13.74 |
| % False negatives | 48.78 | 5.93 | 7.74 | 31.71 | 6.53 | 10.07 |
| Extraction density | 0.49 | 0.99 | 0.88 | 0.69 | 1.80 | 0.93 |
| % Top fields overlap | 62.2 | 60.0 | 64.4 | 55.6 | 62.2 | 62.2 |
| % Date values | 37.5 | 32.1 | 28.3 | 40.2 | 37.4 | 35.8 |
| Fragment length | 34.92 | 41.44 | 46.51 | 57.58 | 28.72 | 48.74 |

Table 4: Evaluation results from different experiments, showing average scores across all input documents.

reduces scores despite factual correctness and highlights that evaluation performance is partly constrained by the robustness of the embedding model. Future work could mitigate such cases by incorporating external norm data or taxonomies for entities such as places and institutions.

The manual analysis further reveals that truly invalid extractions, e.g. units with uninterpretable strings, were rare and were assigned zero scores by the evaluation pipeline. Similarly infrequent were extractions with missing values, which typically reflected underspecified information in the source text. Finally, a small number of FNs received non-zero scores due to alignment with semantically related but incorrect gold fields. Although suboptimal, analysis suggests that these cases had a limited practical and numerical impact on overall performance.

## 7 Conclusion

This study presents, to the best of our knowledge, the first systematic investigation of large-scale IE from biographical texts driven by midsized general-purpose LLMs. The results indicate that the proposed GenIE framework has great potential for effective application in this domain, enabling the extraction of structured biographical information with minimal preparation effort required.

Unlike many similar studies, we adopt an open IE setting that allows for unrestricted output generation. To tailor to this task formulation, we propose a semantic similarity–based evaluation metric. Experiments were conducted on a curated, unpublished dataset of German biographical encyclopedia entries, with domain expert annotations serving as gold references.

We observe that the precision of the IE framework remained relatively stable (approximately 0.61–0.71), while recall varied substantially (approximately 0.33–0.75). While limited precision is partly attributable to incomplete annotation coverage and is therefore less concerning, the absence of high recall remains a notable limitation. At the same time, recall and F1 show a strong positive correlation with extraction density, suggesting that under-generation is more detrimental than moderate over-generation in this setting.

Manual analysis further confirms that many apparent FPs reflect missing gold annotations rather than factual errors, and that the evaluation metric largely behaves as intended. The substantial overlap in extraction field distributions between system output and annotations provides an encouraging signal of structural alignment, despite a systematic bias toward date value extractions.

Controlled ablation experiments reveal that PE strategies offered limited performance gains. Clear prompt formulation and few-shot prompting yield the most consistent improvements, whereas advanced reasoning strategies such as CoT provided little benefit.

Overall, while the output quality of our framework is encouraging, further investigation will be required to fully assess the capabilities of such systems. With appropriate refinement and optimization, the proposed framework may serve as a template for practical deployment. Future research may extend it to new domains, improve open GenIE evaluation methods, and explore a wider range of models.

## Limitations and Ethical Considerations

The dataset used in this study is focused on a specific historical period, namely authors from

the GDR and the evaluation subset primarily features one text type, i.e. encyclopedia entries. Results may not generalize to other domains or (low-resource) languages. Due to resource constraints, experiments used only two open-source LLMs and modest inference parameters. Proprietary models or extensive parameter tuning might yield better performance. This study does not distinguish or quantify different types of hallucinations, focusing instead on the overall occurrence of extraction errors. Evaluation methods for open-ended IE remain imperfect and future research should explore more flexible metrics. Future studies using our proposed metric must ensure the robustness of the core embedding model. Furthermore, the evaluation metric is not yet sensitive to multiple occurrences of events of the same type, as it matches extraction units against all events of the detected type in the annotations for a given source text.

Ethically, LLM use entails significant environmental costs, potential bias, and limited interpretability. Biographical data on living individuals should be extracted only with caution, as it can enable the generation of detailed personal profiles (Ranjan et al., 2022). While such information may serve legitimate purposes, e.g. in law enforcement, it may also carry risks of misuse for malicious purposes.

## Acknowledgements

We are deeply grateful to the FLFDDR project team for creating the corpus that made this study possible. We would also like to thank the reviewers for their valuable and detailed feedback.

## References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie C. Y. Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal M. P. Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Pierre Bourdieu. 1992. *Les règles de l'art: genèse et structure du champ littéraire*. Editions du Seuil, Paris.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health, ML4H@NeurIPS 2023, 10 December 2023, New Orleans, Louisiana, USA*, volume 225 of *Proceedings of Machine Learning Research*, pages 82–100. PMLR.

Bowen Gu, Vivian Shao, Ziqian Liao, Valentina Carducci, Santiago Romero Brufau, Jie Yang, and Rishi J. Desai. 2025. Scalable information extraction from free text electronic health records using large language models. *BMC Medical Research Methodology*, 25(1):23.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *J. Am. Medical Informatics Assoc.*, 31(9):1812–1820.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. 2022. Genie: Generative information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4626–4643. Association for Computational Linguistics.

Muhammad Ali Khan, Umair Ayub, Syed Arsalan Ahmed Naqvi, Kaneez Zahra Rubab Khakwani, Zaryab bin Riaz Sipra, Ammad Raina, Sihan Zhou, Huan He, Amir Saeidi, Bashar Hasan, Robert Bryan Rumble, Danielle S. Bitterman, Jeremy L. Warner, Jia Zou, Amye J. Tevaarwerk, Konstantinos Leventakos, Kenneth L. Kehl, Jeanne M. Palmer, Mohammad Hassan Murad, and 2 others. 2025. Collaborative large language models for automated data extraction in living systematic reviews. *J. Am. Medical Informatics Assoc.*, 32(4):638–647.

Yaxuan Kong, Yuqi Nie, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. Large language models for financial and investment management: Applications and benchmarks. *Journal of Portfolio Management*, 51(2):162 – 210. Cited by: 2.

Jörn Kreutel, Thomas Möbius, Birgit Dahlke, and Steffen Martus. 2023. Forschungsplattform Literarisches Feld DDR – Ein Werkstattbericht zur prosopographischen Erfassung von Schriftsteller:innen in der DDR. In Helmuth Albrecht, Michael Farrenkopf, Helmut Maier, and Torsten Meyer, editors, *Historische Biographik und kritische Prosopographie als Instrumente in den Geschichtswissenschaften*, pages 141 – 165. De Gruyter.

Huaxia Li, Haoyun Gao, Chengzhang Wu, and Miklos A. Vasarhelyi. 2025. Extracting financial data from unstructured sources: Leveraging large language models. *Journal of Information Systems*, 39(1):135 – 156. Cited by: 5.

Shiye Li and Li Yi. 2024. A few-shot entity relation extraction method in the legal domain based on large language models. In *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, DEAI '24, page 580–586, New York, NY, USA. Association for Computing Machinery.

Yu Liu, Duantengchuan Li, Kaili Wang, Zhuoran Xiong, Fobo Shi, Jian Wang, Bing Li, and Bo Hang. 2024. Are llms good at structured outputs? A benchmark for evaluating structured output capabilities in llms. *Inf. Process. Manag.*, 61(5):103809.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self- feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Andrew McCallum. 2005. Information extraction: distilling structured data from unstructured text. *ACM Queue*, 3(9):48–57.

Alistair Plum, Marcos Zampieri, Constantin Orasan, Eveline Wandl-Vogt, and Ruslan Mitkov. 2019. Large-scale data harvesting for biographical data. In *Proceedings of the Third Conference on Biographical Data in a Digital World 2019, Varna, Bulgaria, September 5-6, 2019*, volume 3152 of *CEUR Workshop Proceedings*, pages 66–72. CEUR-WS.org.

Maciej P. Polak and Dane Morgan. 2023. Extracting accurate materials data from research papers with conversational language models and prompt engineering - example of chatgpt. *CoRR*, abs/2303.05352.

Rishabh Ranjan, H. Vathsala, and Shashidhar G. Koolagudi. 2022. Profile generation from web sources: an information extraction system. *Soc. Netw. Anal. Min.*, 12(1):2.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. Promptner: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12492–12507. Association for Computational Linguistics.

Jouni Tuominen, Eero Hyvönen, and Petri Leskinen. 2017. Bio CRM: A data model for representing biographical data for prosopographical research. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017, Linz, Austria, November 6-7, 2017*, volume 2119 of *CEUR Workshop Proceedings*, pages 59–66. CEUR-WS.org.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xiahua Wei, Naveen Kumar, and Han Zhang. 2025. Addressing bias in generative AI: challenges and research opportunities in information management. *Inf. Manag.*, 62(2):104103.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

# A Appendix

| Entire dataset | | Evaluation data | |
|---|---|---|---|
| Source | #Docs | Source | #Docs |
| Deutsche Biographie | 141 | Deutsche Biographie | 34 |
| Deutsches Literatur-Lexikon | 116 | Deutsches Literatur-Lexikon | 21 |
| Wikipedia | 67 | Schriftsteller der DDR (Boettcher) | 13 |
| Schriftsteller der DDR (Boettcher) | 45 | KuerschnerNekrologe1971-1998 | 6 |
| Kuerschner 2012 | 26 | Chemnitzer Autoren | 6 |

Table 5: Top sources in the entire data set vs. evaluation subset.

| Purpose | Prompt part |
|---|---|
| define system role | You are an information extraction assistant. Output must be valid JSON string with the keys 'SUBJECT', 'EVENT_TYPE', 'ATTRIBUTE', 'VALUE' and the correct values corresponding to these keys. |
| instructions | From the input, extract structured information units that each represent a single biographical event and have these fields: 'SUBJECT', 'EVENT_TYPE', 'ATTRIBUTE', 'VALUE'" |
| avoid inference | Only extract information present in the INPUT. Do not add new information or any information from your training data. |
| restrict scope (fields) | Consider only the following EVENT_TYPEs and ATTRIBUTES, skip all others: [...] |
| restrict scope (subject) | Only structured information on the following target person is relevant: [...] |
| source fidelity | Keep the original wording of the short text fragment where the extracted information was found. Do not change the original wording |
| output format | In your response, do not add any explanation or comments, instead return only the structured data. |
| completeness | It is important that you exract as much information as possible!! |
| Few shot instructions | The following is an example for the correct OUTPUT you should generate for a given INPUT. INPUT: |
| example input | INPUT: Görtz, Adolf Erzählung (Kinder- und Jugendliteratur) 31.5. 1920, Köln. 1958-61 Studium am Institut für Lit. „Johannes R. Becher“. Ehrenurk. im Preisausschreiben anläßlich der IV. Weltfestspiele der Jugend und Studenten (1953); 3. Preis im Preisausschreiben des Ministeriums für Kultur (Kinder- und Jugendliteratur) (1953); Lobende Anerkennung im gleichen Preisausschreiben (1954), Leipzig C 1, Jacobstraße 14. Schriftsteller der Deutschen Demokratischen Republik. Bearbeiter des biographischen Teils: Joachim Ret, Egon Sartorius. Bearbeiter des bibliographischen Teils: Helmut Donner, Hans Heininger. Leipzig 1961. |
| example output | fragment_text:“Görtz, Adolf Erzählung (Kinder- und Jugendliteratur)”<br>extracted_info:[SUBJECT:Adolf Görtz,<br>EVENT_TYPE:Employment,<br>ATTRIBUTE:professionsOfEmployment,<br>VALUE:Kinderliteratur, Jugendliteratur]<br><br>fragment_text:“31.5. 1920, Köln”<br>extracted_info:[SUBJECT:Adolf Görtz,<br>EVENT_TYPE:Birth,<br>ATTRIBUTE:begin,<br>VALUE:1920-05-31,<br>SUBJECT:Adolf Görtz,<br>EVENT_TYPE:Birth,<br>ATTRIBUTE:place,<br>VALUE:Köln]<br><br>fragment_text:“1958–61 Studium am Institut für Lit. “Johannes R. Becher” ”<br>extracted_info:[SUBJECT:Adolf Görtz,<br>EVENT_TYPE:Education,<br>ATTRIBUTE:institution,<br>VALUE:Literaturinstitut Johannes R. Becher,<br>SUBJECT:Adolf Görtz,<br>EVENT_TYPE:Education,<br>ATTRIBUTE:begin,<br>VALUE:1958,<br>SUBJECT:Adolf Görtz,<br>EVENT_TYPE:Education,<br>ATTRIBUTE:end,<br>VALUE:1961]<br><br>fragment_text:“Ehrenurk. im Preisausschreiben anläßlich der IV. Weltfestspiele der Jugend und Studenten (1953)”<br>extracted_info:[SUBJECT:Adolf Görtz,<br>EVENT_TYPE:<br>[...] |

Table 6: Prompt used in the experiments (shortened and re-formatted for readability).