Information Processing and Management 63 (2026) 104833

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Synthetic dataset generation for theft event extraction in Italian

Giovanni Bonisoli *, Federica Rollo, Laura Po

"Enzo Ferrari" Engineering Department, University of Modena and Reggio Emilia, Via Vivarelli, 10 - 41125 Modena, Italy

ARTICLE INFO

Dataset link: https://github.com/federicarollo/SYNTH-ITA

Keywords:
Synthetic
Event-extraction
Italian
Crime
LLMs
Validation

ABSTRACT

Event extraction is the task of automatically identifying and extracting structured information about events from unstructured text. Despite Italian being a well-resourced language, it still lacks annotated datasets specifically designed for fine-grained event extraction. To address this gap, we propose a novel methodology for the generation of synthetic data suitable for fine-grained event extraction tasks. This work is motivated by the high cost and limited scalability of manual annotation. We introduce a controlled synthetic data generation pipeline that strictly adheres to a target annotation schema, providing a scalable alternative to extensive human labeling. The key methodological innovation is a two-phase, document-level generation framework that leverages Large Language Models, ensures structural consistency and mitigates generation biases, enabling the creation of high-quality datasets for complex event extraction scenarios.

Using this methodology, we release SYNTH-ITA, the first collection of four medium-scale synthetic datasets for fine-grained Italian event extraction, generated from 10,000 structured crime scenarios each. Experiments conducted on event argument extraction using a QA formulation demonstrate that fine-tuning models on SYNTH-ITA leads to better or comparable performances to models fine-tuned on 200 manually annotated real news articles (+14% improvement with ELECTRA, -0.4% with BERT). Conversely, NER-based models for event argument extraction trained on synthetic data exhibit an 18% performance drop compared to those trained on manually annotated articles.

1. Introduction

Data augmentation is a key strategy in Natural Language Processing (NLP) to increase the size and diversity of training data, especially when annotated resources are scarce. It involves generating new data from existing examples or guidelines to enhance model generalization and robustness. Recent surveys highlight its impact on a wide range of NLP tasks, from classification to sequence labeling and generation (Chen et al., 2023; Feng et al., 2021; B. Li et al., 2022).

In recent years, the advent of generative AI has further expanded data augmentation capabilities. Techniques such as Variational Autoencoders (VAEs) (Kingma & Welling, 2022), Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), and more recently diffusion models (Ho et al., 2020) have been applied to produce synthetic data that improves the generalization ability of discriminative models, particularly in the text domain de Rosa and Papa (2021), Piedboeuf and Langlais (2022), Yi et al. (2024). The emergence of Large Language Models (LLMs) has marked a turning point, offering the ability to produce high-quality, controllable text from structured prompts. This capability enables the construction of large-scale synthetic datasets that can enrich or even surpass human-curated corpora for specific downstream NLP tasks (Ding et al., 2024).

* Corresponding author.
E-mail addresses: giovanni.bonisoli@unimore.it (G. Bonisoli), federica.rollo@unimore.it (F. Rollo), laura.po@unimore.it (L. Po).

https://doi.org/10.1016/j.ipm.2026.104833
Received 9 August 2025; Received in revised form 20 March 2026; Accepted 15 April 2026
Available online 6 May 2026
0306-4573/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

This work is driven by two critical gaps in current event extraction research: (1) the scarcity of annotated resources for fine-grained event extraction in Italian, and (2) the lack of scalable methods for generating training data at the document level. Most existing synthetic data generation techniques for event extraction operate at the sentence level (Tian et al., 2024; Wang & Huang, 2024), assuming all event information is contained within single sentences. This limitation prevents their application to real-world scenarios involving complex events distributed across multiple sentences, coreference chains, or multiple mentions per role. In low-resource settings, traditional methods for event extraction are often unsuitable due to their reliance on large amounts of annotated training data. Bonisoli et al. (2025) show that instruction-tuned LLMs can offer a viable alternative with minimal annotated data, however the results remain modest, with LLMs achieving an F1-score of around 60%, highlighting both their potential and the need of a different approach for further improvement.

To address these gaps, we introduce a document-level synthetic data generation pipeline with an explicit alignment mechanism that ensures structural fidelity between complex event schemas and generated multi-sentence narratives. Unlike sentence-level approaches that rely on paraphrasing existing text, our method: (1) generates documents from structured schemas rather than reference instances, (2) employs rule-based post-processing for span-level verification and correction, and (3) incorporates expert feedback in an iterative refinement loop specifically targeting document-level coherence issues. Our ultimate goal is to train event extraction models that outperform those trained solely on the limited amount of available real annotated data.

1.1. Research objectives

The main goal of this research is to propose a methodology for generating high-quality, diverse, and controllable synthetic data using LLMs. We introduce SYNTH-ITA, a collection of four medium-scale synthetic datasets, each containing thousands of high-fidelity Italian theft news articles paired with fine-grained event annotations. These annotations, automatically generated and validated, capture critical event details such as perpetrators, victims, locations, and stolen objects, providing a substantial resource for training and evaluating information extraction systems.

Our research objectives are threefold:

- Design and implementation of a controlled, document-level LLM-based pipeline for generating synthetic documents. Unlike sentence-level approaches, our methodology is specifically designed to handle multi-sentence narratives with complex event structures, starting from structured annotations.
- Construction of medium-scale datasets of automatically generated theft-related news articles, each enriched with fine-grained, structured event annotations, ensuring semantic fidelity and fairness safeguards.
- Evaluation of the synthetic data effectiveness by training and testing Italian NLP models (for tasks such as Named Entity Recognition and Question Answering) and comparing their performance with models trained on real annotated datasets.

1.2. Contributions

SYNTH-ITA offers a substantial advancement over existing resources by introducing, for the first time, synthetic datasets for fine-grained Italian event extraction that combine scale, granular, event-centric annotations, and bias mitigation. Although it is based on the annotation schema of the DICE dataset (Bonisoli et al., 2023), SYNTH-ITA is entirely new in content, significantly larger in scale, and explicitly tailored for structured event analysis in the Italian language.

Our main contributions are:

1. A controlled, two-phase, document-level synthetic data generation pipeline with an explicit alignment mechanism that combines schema-driven generation, rule-based post-processing, and iterative expert validation to maintain structural fidelity in multi-sentence event narratives.
2. The first set of public available medium-scale synthetic datasets for fine-grained Italian event extraction, comprising four datasets generated from 10,000 structured theft-related annotations.
3. Integration of fairness safeguards, ensuring balanced distributions of gender and nationality in the generated data.
4. Expert validation for high-quality synthetic datasets and extensive automatic evaluation on textual quality, syntactic diversity, semantic alignment and annotation faithfulness.
5. Italian resources for downstream NLP tasks (Named Entity Recognition and Question Answering formulations for event extraction).

1.3. Outline

The remainder of this paper is organized as follows. Section 2 provides an overview of existing Italian datasets for event extraction and related work on data augmentation techniques and synthetic datasets for NLP, with a particular focus on event extraction. The methodology adopted to create the SYNTH-ITA datasets, outlining the two-phase process involving the validation and selection of LLMs for realistic news generation and the synthetic dataset generation, is depicted in Section 3. Section 4 describes the application of the methodology to a specific case study, i.e., the Italian crime news, and presents the automatic and manual evaluation of the generated datasets. Section 5 discusses the use of SYNTH-ITA datasets for Named Entity Recognition and Question Answering formulations for event extraction. Section 7 highlights some practical implications and limitations of this work, while Section 8 sketches conclusion and suggests future directions.

2. Related work

2.1. Italian datasets for event extraction

Research on event extraction in Italian is constrained by the limited availability of datasets with fine-grained event annotations. Existing corpora either lack explicit event-level information, provide only temporal or entity annotations, or include detailed annotations for a very restricted set of events. As a result, none of the current resources is sufficiently detailed to train event extraction models. The only exception is DICE (Bonisoli et al., 2023), which offers fine-grained manual annotations, but the annotated portion is too small (606 news articles) to support robust model training. This motivates the creation of a large-scale, synthetically annotated dataset for fine-grained Italian event extraction.

Large textual resources such as La Repubblica[1] (Baroni et al., 2004) and PAISÀ[2] (Lyding et al., 2014) provide valuable linguistic material, including POS tags, but they do not include event-centric annotations. Their primary purpose is general linguistic analysis, not structured event modeling.

The EVENTI corpus[3] focuses on temporal information processing, including EVENT, TIMEX3, SIGNAL and TLINK annotations. However, its event tagset is not designed to capture event roles, participants, or multi-span structures. Similarly, MEANTIME (Minard et al., 2016) provides multilingual annotations of entities, events, and temporal relations, but the Italian portion remains relatively small (120 documents). The De Gasperi corpus (Tonelli et al., 2019) provides metadata and named-entity annotations over historical documents but only references to persons and places are annotated.

The only Italian dataset offering fine-grained event annotations is DICE (Bonisoli et al., 2023), a corpus of crime-related news articles. While the full collection includes over 10,000 documents, only 606 theft-related articles are manually annotated with a detailed multi-span schema (including What, Where, Who and associated socio-demographic attributes). Despite its high annotation quality, the extremely limited number of fully annotated texts prevents its use as the sole training resource for event extraction architectures, which typically require much larger datasets.

More recently, EventNet-ITA[4] (Rovera, 2024) has introduced large-scale frame parsing annotations for Italian. Although frames provide valuable semantic structure, they do not correspond directly to fine-grained event extraction schemas: roles are defined at the frame level without considering co-reference.

2.2. Data augmentation for event extraction

Data augmentation is vital in event extraction due to the high cost of manual annotation. Gao et al. (2022) developed Mask-then-Fill, a framework that flexibly edits text to generate diverse data while preserving event structure, addressing the limitations of synonym replacement, back-translation, and BERT-based methods. Wang et al. (2023) proposed DAEE, a denoised structure-to-text augmentation framework that employs a deep reinforcement learning agent to address key limitations of data augmentation methods, including structure misalignment, grammatical errors, and semantic drift. Subsequent studies have proposed LLM-based augmentation comprising a generation step followed by a verification step. For instance, TALOR-EE (Wang & Huang, 2024) prompts an LLM with a structured event template to produce fluent, semantically accurate sentences that correspond to a specific event type, and then employs an NLI-based entailment module to verify their coherence with the template. Similarly, AGENT-DA (Tian et al., 2024) adopts a collaborative multi-agent framework to achieve the same objective. All these methods target sentence-level event extraction, which assumes that all event information is contained within a single sentence. Moreover, in these methods, the generated text is consistently derived from a reference instance and preserves the same event annotations as the original text. Such a reference-driven augmentation strategy may introduce systematic biases, as models trained on the resulting synthetic data and evaluated on real-world texts with identical annotations may overfit to annotation-specific patterns rather than learning robust event representations, thereby limiting their generalization capability.

At the document level, Jin and Ji (2024) presented a schema-based approach using schema graphs to represent domain-specific events and relations, sampling subgraphs prioritizing frequent nodes, populating them with real-world examples, and converting them into text via generative models like GPT-3.5. Other works have examined the subtask of document-level Event Argument Extraction (EAE). Liu et al. (2021) propose two methods: implicit knowledge transfer using reformatted labeled data from related tasks within a machine reading comprehension (MRC) framework, and a pre-trained MRC model that generates labeled examples from unlabeled data. They later introduced a noise filtering strategy to enhance data quality by selecting reliable instances (Liu et al., 2022). Most recently, Gatto et al. (2025) explored the use of LLMs to generate synthetic samples for EAE, with the aim of addressing real-world challenges such as processing long documents and handling event roles with limited training data. A common limitation of these approaches is their inability to handle multi-span event roles and document-level event structures, in contrast to our method. Most existing synthetic data generation techniques for event extraction operate at the sentence level and rely on reference-driven augmentation, where new samples are derived from existing annotated instances while preserving their original event structure. This assumption limits their applicability to scenarios involving complex events distributed across multiple sentences, coreference chains, or multiple mentions per role. Moreover, these methods typically focus on increasing lexical or syntactic diversity, without explicitly controlling the global consistency between structured event representations and generated text. For these reasons, such approaches are not directly comparable to the document-level, schema-driven generation framework proposed in this work.

[1] https://docs.sslmit.unibo.it/doku.php?id=corpora:repubblica
[2] https://www.corpusitaliano.it
[3] https://sites.google.com/site/eventievalita2014/data-tools
[4] https://huggingface.co/datasets/mrovera/eventnet-ita

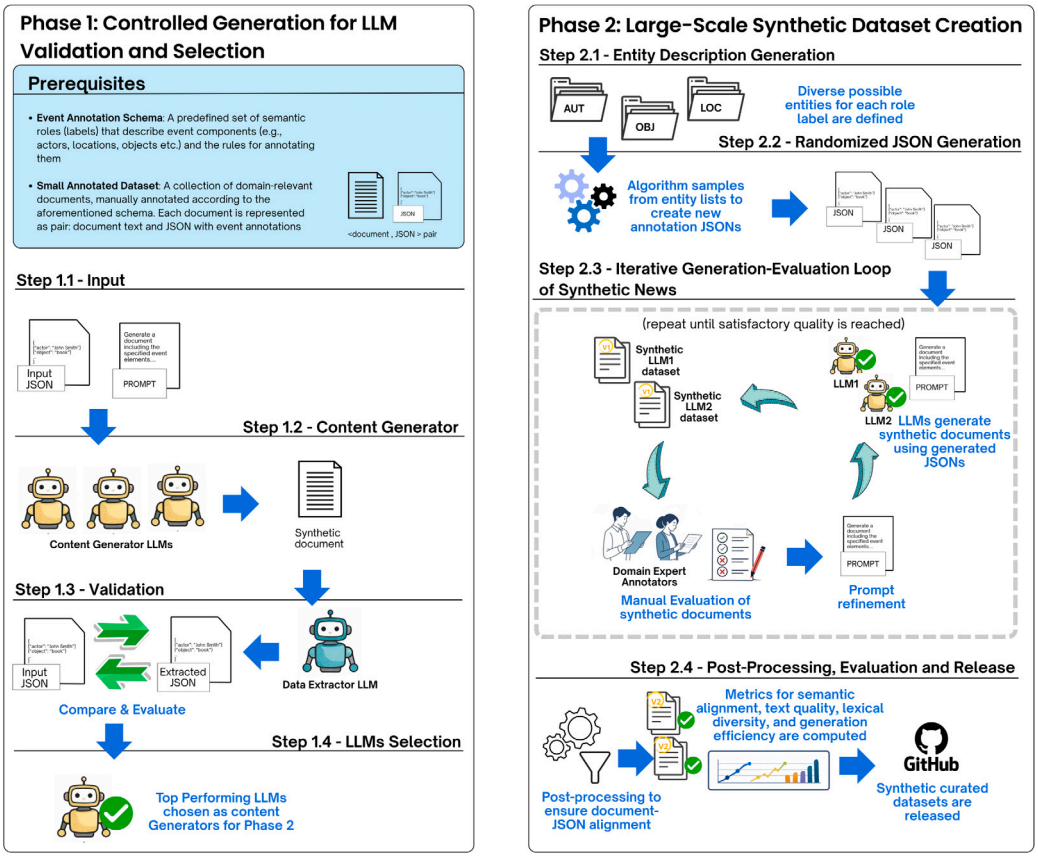

Fig. 1. Workflow of Phase 1 and 2.

2.3. Synthetic dataset generation with LLMs

Some works in literature demonstrated that LLMs can generate synthetic training data even for tasks with structured outputs, by prompting the model to generate plausible input text from a target structured output. Josifoski et al. (2023) proposed SynthIE, an approach that leverages triplets sampled from a knowledge graph as input to LLMs for generating high-quality synthetic data. Ye et al. (2023) developed SymGen, which improves symbolic language data generation using informative prompts and verification strategies, enabling efficient training of smaller models. Similar strategies have been applied for text-pair classification tasks (Li et al., 2025), Named Entity Recognition (Ye et al., 2024) and automatic dialogue evaluation (Ye et al., 2025). In literature, examples of LLMs-generated dataset are in the domain of Information Retrieval (Bonifacio et al., 2022) and reasoning (Liu et al., 2023). Some data augmentation frameworks are designed to operate across multiple NLP tasks. Among these, AugGPT (Dai et al., 2025) rephrases each sentence into several semantically similar variants of the original data using ChatGPT. However, simple paraphrasing often fails to preserve the nuanced event-role structures required for Event Extraction, highlighting the need for structure-aware augmentation methods. Another example is Self-LLMDA (Li et al., 2024), which reduces human effort by using LLMs also for the generation of augmentation instructions and applying a scoring model based on FLAN-T5 (Chung et al., 2024) for the selection of task-specific instructions. Currently, Self-LLMDA is mainly focused on English language and not easily adaptable to Italian.

Challenges remain in ensuring quality and fairness, as Nadăş et al. (2025) identified risks like factual errors and bias amplification, suggesting mitigation techniques such as filtering and reinforcement learning. Differently, our approach explicitly incorporates bias mitigation at the input generation level and includes expert review to detect and correct systemic issues in the output. Some approaches leverage LLMs for data-to-text generation by transforming structured input into human-readable descriptions. For example, Puduppully and Lapata (2021) developed a methodology similar to our two-step methodology involving content planning followed by text generation, aiming to improve the faithfulness between structured input and generated text. However, these works target a different tasks (e.g., sports reporting, information retrieval) and are conducted in English, whereas our focus is on news generation in Italian. Feng et al. (2025) leverage LLMs to generate synthetic event pairs for representation learning, focusing on semantic equivalence. Differently, SYNTH-ITA targets the creation of fully annotated event extraction datasets, ensuring structural fidelity through a controlled generation pipeline.

3. Methodology

3.1. Overview and purpose

The proposed methodology aims to generate high-quality synthetic datasets for event extraction, starting from a limited set of manually annotated real documents. The core concept leverages a structured event annotation schema to guide LLMs in a controlled generation of realistic texts. These texts, semantically aligned with the target event schema, constitute a large-scale synthetic dataset. The ultimate goal is to train event extraction models that outperform those trained solely on the limited amount of available real annotated data.

Methodological Innovation. While individual components of our pipeline (LLM prompting, structured generation) build upon established techniques, the novelty of our work lies in their integration into a controlled, document-level, schema-driven generation framework with an explicit alignment mechanism that ensures structural fidelity between complex event schemas and generated multi-sentence narratives. Unlike existing sentence-level or reference-driven augmentation methods for event extraction (Tian et al., 2024; Wang & Huang, 2024), our approach is specifically designed to generate coherent, multi-sentence narratives that adhere to complex event schemas involving coreference and multi-span arguments. This addresses a key limitation in the field: the lack of scalable methods for creating training data for document-level event extraction.

This section focuses on describing the controlled pipeline for synthetic event dataset generation. The overall schema of the pipeline is illustrated in Fig. 1.

3.2. Prerequisites: Event schema and real annotated data

The method assumes the existence of two initial resources:

1. **An event annotation schema**: A predefined set of semantic roles (*labels*) that describe the key components of an event (e.g., actors, locations, objects etc.) and the rules for annotating them, including complex cases such as multi-word mentions, coreference, and multiple assignments.
2. **A small real annotated dataset** (typically 100–200 documents): A collection of domain-relevant documents, manually annotated according to the aforementioned schema, serving both as a reference for linguistic fidelity and as a basis for validation. Each document is represented as paired `(document_text, annotation_JSON)`, where the JSON structures the extracted information according to the event annotation schema's roles.

3.3. Two-phase controlled generation pipeline

The methodology unfolds in two sequential phases, designed to ensure control, quality, and scalability.

3.3.1. Phase 1: controlled generation for LLM validation and selection

The objective of this phase is to validate and select the most suitable LLMs for content generation by assessing their ability to produce documents that faithfully adhere to structured input annotations. Using the real `(document, JSON)` pairs as a reference, we evaluate multiple candidate LLMs based on comprehensive metrics measuring semantic fidelity, textual quality, and structural alignment.

1. **Step 1.1 - Input**: The set of real JSONs derived from manual annotations of the small real dataset.
2. **Step 1.2 - Content Generation**: Each candidate LLM is prompted with specific instructions and a JSON to generate a document that incorporates *all* specified fields while maintaining a linguistic style consistent with the domain (e.g., journalistic). Multiple prompting strategies (zero-shot, few-shot) are evaluated.
3. **Step 1.3 - Validation**: A pre-validated *Data Extractor* LLM processes each generated synthetic text to extract a new JSON. The extracted JSONs are compared with the input JSONs to evaluate the structural alignment. Moreover, multiple automated metrics for semantic alignment, text quality, lexical diversity and generation efficiency are computed on the generated documents.
4. **Step 1.4 - LLMs selection**: The candidate LLMs demonstrating the best balance across these metrics are selected as the *Content Generator* LLMs for Phase 2.

3.3.2. Phase 2: Large-scale synthetic dataset creation

This phase builds the final synthetic dataset through an iterative process that incorporates manual expert evaluation and prompt refinement to ensure high quality and realism.

1. **Step 2.1 - Entity Description Generation**: For each semantic role in the annotation schema, lists of possible entity descriptions are defined (e.g., for the "Location" role: a list of realistic places). These lists can be extracted from the real dataset or manually curated to ensure diversity and realism.
2. **Step 2.2 - Randomized JSON Generation**: An algorithmic process pseudo-randomly samples from these lists to create new JSON annotations. The process can incorporate distributional constraints (e.g., frequencies of certain roles) and fairness safeguards (e.g., balanced gender or nationality).

3. **Step 2.3 - Iterative Generation-Evaluation Loop of Synthetic News** Using the generated JSONs as input, the LLM produces new documents through a predefined generation prompt (mirroring Step 1.2). A critical quality-control step involves domain experts manually evaluating a random sample of the initial synthetic dataset. This expert review served multiple purposes: (1) to verify the synthetic data accurately represents real-world scenarios, (2) to identify any biases that might have been unintentionally introduced during generation, and (3) to confirm the synthetic data generated meets established annotation guidelines. Experts annotate the sample, and their annotations are compared against the original JSONs. They also identify recurring issues (e.g., grammatical errors, inconsistent information, or label misinterpretations). Based on this analysis, the generation prompt is systematically refined to address the identified shortcomings. This process may be repeated until predefined quality thresholds are met. Once the prompt refinement process is completed, the final prompt is used to generate the full synthetic dataset at scale.
4. **Step 2.4 - Post-Processing, Evaluation and Released**: A rule-based post-processing algorithm (Section 3.4) is applied to align the generated text with its corresponding JSON annotations. Items failing the alignment checks are filtered out, resulting in a final high-fidelity synthetic corpus. Automated metrics for semantic alignment, text quality, lexical diversity, and generation efficiency are computed on the post-processed documents to assess the overall quality of the synthetic dataset. The resulting validated dataset is then released as a curated synthetic resource for downstream research and evaluation.

3.4. Post-processing for explicit JSON-text alignment

To ensure a reliable synthetic dataset where each generated text faithfully contains the information specified in its structured annotation, we introduce a rule-based post-processing algorithm that performs **explicit, span-by-span alignment verification and correction of the structured JSON**. This mechanism is crucial for two reasons: (1) it guarantees that the released dataset consists of high-fidelity `(document, JSON)` pairs, and (2) it ensures that models trained on the dataset for downstream event extraction can find the target arguments within the text, as the alignment is verified and enforced.

The algorithm operates through a sequence of targeted operations applied to each generated news article and its corresponding input JSON:

1. **Span Verification**: Each annotation span in the JSON is checked for its presence in the generated text.
2. **Alternative Formulation Search**: If a span is not found, the system searches the text for semantically equivalent mentions using rule-based pattern matching:
    - **Attribute Reformulation**: Detects alternative textual formulations of the same attribute (e.g., *"a 38-year-old man"* ↔ *"38 years old"*; *"of Italian nationality"* ↔ *"Italian"*).
    - **Morpho-syntactic Adjustment**: Harmonizes number, case, or determiner variations (e.g., singular ↔ plural form, business name capitalization).
    - **Synonym Replacement**: Identifies synonyms or related terms (e.g., *"cell phone"* ↔ *"mobile phone"*).
3. **JSON Update**: When an equivalent formulation is identified, the original JSON span is updated to match the exact textual mention, ensuring perfect alignment for downstream extraction.
4. **Omission Handling**: Spans for which no alternative formulation is found are flagged as omitted. A configurable policy then determines the final outcome:
    - **Span Removal**: The omitted span is removed from the JSON annotation if the information is deemed non-essential.
    - **Item Filtering**: The entire news item is discarded if the omission concerns a critical, non-recoverable event argument.

    The identification of non-essential or critical information depends on the case study.

This post-processing step is not a simple cleanup but a core component of our **controlled generation pipeline**. By reconciling the inevitable variations between LLM output and structured input, it maximizes the yield of usable `(text, JSON)` pairs while preserving semantic fidelity. The result is a synthetic dataset where the alignment between unstructured narratives and structured event records is explicitly verified and corrected, making it a reliable resource for training and evaluating document-level event extraction models.

3.5. Implementation with large language models

The choice of the specific Content Generator LLMs is empirically driven by the comprehensive evaluation in Phase 1, which balances semantic alignment, textual quality, and operational efficiency. The subsequent **iterative prompt refinement** in Phase 2, informed by manual expert evaluation, further optimizes the generator's output, ensuring it meets the stringent coherence and alignment standards required for our document-level event extraction benchmark. This model-agnostic yet rigorously validated design ensures the pipeline's robustness and adaptability.

Table 1
Summary of evaluation metrics.

		Metrics	What Is Measured	Why Is Measured	Phase 1	Phase 2
Intrinsic quality and realism	Semantic Alignment	*Between synthetic and real documents sharing the same annotation*: BERTScore (BS) *Within document*: Sentence-BERT (**sBS**)	Preservation of meaning between real and synthetic text sharing the same annotations and internal coherence	Ensures that generated text is unambiguous and internally coherent: EE models cannot be trained on text that contradicts itself or leaves events ambiguously described, as this would prevent learning consistent annotation patterns	✓	
	Textual Quality	*Within document*: Document length (**len**); Avg sentence length ($\textbf{len}_{\textbf{sen}}$); Vocabulary size (**voc**); Part-of-Speech distribution (POS); Gulpease Index (GI); Perplexity (PLX)	Fluency, readability, and linguistic naturalness	Ensures that models learn from natural linguistic patterns: training on disfluent or unnatural text causes models to memorize generation artifacts rather than generalize to real-world event expressions	✓	✓
	Lexical Diversity	*Within document*: Dist-N, MTLD, HD-D, MATTR *Across documents of the same dataset*: Div-N, Self-BLEU, Self-repetition score *Between a dataset and a reference dataset*: JSD-N	Lexical and structural variability across and within documents	Prevents overfitting to narrow patterns, low diversity yields brittle models that fail on unseen event formulations	✓	✓
	Structural Alignment	*Between a document and the corresponding JSON*: Exact Match (EM); Partial Match (PM); Omitted spans (**omit**); Items with zero omitted spans ($\#\text{ne}_{\text{no-omit}}$)	Faithfulness of synthetic text to input annotation schema	Directly determines if synthetic data are useful for training EE models: training is possible only on annotations that appear in the generated text, omitted spans cannot contribute to learning argument extraction	✓	✓
	Operational Efficiency	*Within dataset*: Generation time/item (**gen**); Acceptance rate ($\#\text{ne}_{\text{released}}/\#\text{ne}_{\text{generated}}$)	Practical scalability and efficiency	Determines practical viability: high generation speed is insufficient if most outputs are invalid; true efficiency requires balancing throughput with yield	✓	✓

3.6. Evaluation framework and metrics

Evaluating synthetic datasets presents a dual challenge: ensuring linguistic realism while maintaining structural adherence to the input schema. Manual qualitative analysis is feasible only for small samples, but becomes unscalable for medium-scale and large-scale datasets. We therefore adopt a multi-dimensional automated framework to benchmark synthetic data against real-world references and quantify the reliability of our generation pipeline.

Table 1 summarizes our evaluation framework. Metrics are grouped by evaluation dimension, each targeting specific quality aspects, and are designed to be applicable across different event types and domains. The table also indicates in which experimental phase each metric is applied with different scopes:

- **Phase 1 (LLM Validation)** aims to select the most suitable LLMs for content generation. It combines metrics that compare generated documents to the small real annotated dataset with metrics assessing intrinsic quality and generation efficiency. The selected LLMs are those that best balance semantic and structural alignment, textual quality, diversity, and operational efficiency.
- **Phase 2 (Dataset Quality Assessment)** shifts focus to evaluating the intrinsic quality of the novel synthetic dataset. This phase applies all textual quality, diversity, and structural alignment metrics to the generated corpus. Crucially, it also incorporates expert manual evaluation to identify issues and enable iterative prompt refinement, ensuring the final dataset meets high standards of coherence and realism.

In the following, we present each metric and describe how it is calculated.

BERTScore (BS)[5] (Zhang et al., 2020) is a semantic similarity metric that computes the alignment between two texts via greedy matching of contextual token embeddings. We rescale raw scores using an empirical baseline from random sentence pairs to obtain interpretable $[0, 1]$ values, where scores close to 1 indicate strong semantic alignment.
Sentence-BERT (sBS) computes sentence-level similarity via cosine distance between Sentence-BERT embeddings (Reimers & Gurevych, 2019), producing scores in $[0, 1]$ where higher values indicate greater semantic similarity.
Basic Statistics includes average document length (**len**), average sentence length ($\mathbf{len_{sen}}$), vocabulary size (**voc**), and Part-of-Speech (**POS**) distribution in each document.
Gulpease Index (GI) estimates the Italian readability based on the number of sentences (n_s), the number of letters (n_l) and the number of words (n_w): $89 + (300 \cdot n_s - 10 \cdot n_l) \cdot n_w^{-1}$. Scores below 80 indicate texts difficult for elementary readers; below 60, difficult for middle school; below 40, difficult for high school level.
Perplexity (PLX) (Chen & Goodman, 1996) measures how well a language model predicts a text. Its values range from 1 to infinity: lower values indicate more fluent and predictable text. We compute it using Minerva-3B-base-v1.0,[6] an Italian model that provides more accurate probability estimates for Italian texts than multilingual alternatives.
Dist-N (Li et al., 2016) is the proportion of unique n-grams within a document $[0, 1]$. Higher values indicate greater within-document lexical diversity.
MTLD (Measure of Textual Lexical Diversity) (McCarthy & Jarvis, 2010) computes average length of sequential word segments maintaining a TTR (Token-Type Ratio) ≥ 0.72. TTR (Association et al., 1944) is the proportion of unique words in the sequence. Higher values indicate sustained lexical diversity over longer text spans.
HD-D (Hypergeometric Distribution Diversity) (McCarthy & Jarvis, 2010) estimates expected distinct tokens in random token samples using hypergeometric distribution. It ranges from 0 to 1, higher values indicate greater vocabulary richness.
MATTR (Moving-Average Type/Token Ratio) (Covington & McFall, 2010) computes TTR over fixed-length sliding windows and averages them. It ranges from 0 to 1, higher values indicate greater lexical variation per window.
Div-N is a normalized metric $[0, 1]$ measuring the proportion of n-grams in a document that do not appear in any other document of the same dataset. Values close to 0 indicate no novelty (all n-grams are shared across documents); values close to 1 indicate maximum novelty (all n-grams are unique to that document).
Self-BLEU (Zhu et al., 2018) measures the corpus-level diversity $[0, 1]$ as the average BLEU score of each text against all others in the same dataset. Values close to 1 indicate low diversity (high inter-document similarity); values near 0 indicate high diversity (low similarity between documents).
Self-repetition score (Salkar et al., 2022) is a corpus-level metric $[0, 1]$ measuring the proportion of documents that share at least one repeated n-gram (length ≥ 4) with another document in the same dataset. A score of 0 indicates no document shares repeated phrasing; a score of 1 indicates every document contains at least one repeated n-gram found elsewhere in the corpus.
Jensen–Shannon Divergence (JSD-N) (Lu et al., 2020) is a symmetric bounded metric $[0, 1]$ that measures the distance between two probability distributions (e.g., n-gram distributions of generated vs. reference texts). Lower values indicate more similar distributions; higher values indicate greater divergence.
Exact Match (EM) (H. Li et al., 2022) evaluates exact span-level correspondence between two sets of annotations. Treating one set as predictions and the other as ground truth, a true positive occurs when a predicted span exactly matches a ground-truth span; false positives are predictions with no exact match; false negatives are ground-truth spans not exactly matched by any prediction. We report micro-averaged precision, recall, and F1-score.
Partial Match (PM) (H. Li et al., 2022) measures partial span overlap via longest common substring. For each prediction, we take its maximum overlap (relative to prediction length) with any ground-truth span; for each ground-truth, we take its maximum overlap (relative to ground-truth length) with any prediction. Precision and recall are the respective averages, yielding scores in $[0, 1]$ where higher values indicate greater partial overlap.
Omitted spans (omit) indicates the percentage of annotation spans present in the input JSON that are missing from the generated text before applying post-processing.
$\mathbf{\#ne_{no\text{-}omit}}$ represents the number of items with zero omitted spans (w.r.t. the input JSON).
gen indicates the average time required to generate one document.
Acceptance rate refers to the proportion of released items ($\mathbf{\#ne_{released}}$) w.r.t. the generated items ($\mathbf{\#ne_{generated}}$).

3.6.1. Phase-specific application

Phase 1 (LLM Validation): The primary goal is to select the most suitable LLMs for content generation. This phase combines metrics that compare generated documents to the small real annotated dataset with metrics assessing intrinsic quality of the generated documents and generation efficiency. The selected LLMs are the one that best balances semantic and structual alignment, textual quality, diversity, and operational efficiency.

Phase 2 (Dataset Quality Assessment): The focus shifts to evaluating the intrinsic quality of the novel synthetic dataset. This involves applying all textual quality, diversity, and structural alignment metrics to the generated corpus. Crucially, this phase incorporates expert manual evaluation to identify issues and enable iterative prompt refinement, ensuring the final dataset meets high standards of coherence and realism.

[5] https://github.com/Tiiiger/bert_score
[6] https://huggingface.co/sapienzanlp/Minerva-3B-base-v1.0

Table 2
Evaluation of Phase 1.

Dataset	#ex	gen	len	Textual Quality		Diversity		JSON faithfulness			Semantic Alignment	
				GI	PLX	Div-2	Div-3	omit	EM	PM	BS	sBS
$DICE_{406}$	–	–	1489	50	9.5	0.97	1.00	–	–	–	–	–
$SYN_{LL3\text{-}8B}$	0	11.1	565	54	13.0	0.81	0.89	44.7	47.3	59.3	16.4	69.8
	1	10.4	529	55	12.4	0.94	0.98	37.9	48.6	60.2	20.3	72.5
	2	11.2	569	56	13.3	0.94	0.98	36.3	51.8	64.9	21.1	74.0
$SYN_{LLT3\text{-}8B}$	0	8.0	380	53	16.4	0.85	0.92	46.2	48.0	59.6	18.3	69.5
	1	8.6	421	51	16.2	0.95	0.99	44.3	50.7	62.4	19.1	71.4
	2	10.0	484	52	18.1	0.95	0.99	37.9	48.2	59.1	18.3	72.7
SYN_{M7B}	0	20.9	590	52	14.7	0.89	0.96	46.3	40.7	52.5	15.5	67.5
	1	15.5	483	53	12.4	0.95	0.99	43.9	49.1	61.6	19.4	71.8
	2	19.5	558	54	11.9	0.95	0.99	40.9	46.8	60.8	19.7	72.5
SYN_{Z7B}	0	10.6	946	51	11.8	0.92	0.97	54.0	42.9	54.0	10.8	66.4
	1	10.7	637	53	13.6	0.96	0.99	47.2	48.7	59.4	16.9	69.6
	2	12.7	796	53	12.7	0.96	0.99	45.1	47.7	60.2	15.9	69.5
SYN_{Q7B}	0	1.8	318	51	13.9	0.91	0.97	28.6	61.8	72.9	12.9	62.8
	1	5.4	1063	48	8.9	0.92	0.98	19.0	58.7	70.6	14.0	73.6
	2	5.9	1140	46	8.9	0.91	0.97	18.6	58.5	71.0	13.1	72.5
SYN_{Q14B}	0	3.3	384	51	14.3	0.93	0.97	18.3	64.7	75.5	13.7	65.6
	1	5.0	582	49	9.8	0.89	0.95	23.3	62.0	75.3	15.0	70.3
	2	6.4	756	47	8.6	0.96	0.99	19.6	62.7	75.6	15.0	72.3
SYN_{DS}	0	111.7	1460	46	6.7	0.87	0.95	17.1	58.0	72.2	11.2	71.1
	1	75.6	897	48	6.5	0.90	0.96	15.3	62.2	76.0	14.1	72.2
	2	83.7	1001	49	5.8	0.91	0.97	15.3	61.6	76.4	15.0	73.3

4. Synthetic dataset generation and evaluation: Case study on Italian crime news

This section describes the application of the general methodology introduced in Section 3 to a specific domain: Italian crime news, with a particular focus on theft events. We adopt the DICE dataset (Bonisoli et al., 2023) as our reference real-world corpus, which comprises 10,395 Italian crime news articles. DICE is currently the only available Italian dataset for fine-grained event extraction that provides a high-quality, validated annotation schema for theft events. In addition, it includes a set of 606 manually annotated theft-related articles, which serves as a reliable benchmark for system development and evaluation.

Our focus on thefts is motivated by both methodological and empirical considerations. The availability of an established annotation schema (only for theft) and a consistent manually annotated subset enables rigorous experimentation, whereas constructing an equally reliable schema for a different event type would require a substantial and time-consuming annotation effort. Moreover, thefts are the predominant event type in DICE, accounting for 73.37% of all annotated instances, which allows us to draw on a large and diverse pool of source articles. Such representativeness and variability are not attainable for less frequent categories, such as drug dealing or assaults.

Our goal is to instantiate the SYNTH-ITA dataset by applying the two-phase pipeline to this domain. To this end, following the same subdivision used by Bonisoli et al. (2025), we split the set of manually annotated news articles into three disjoint subsets consisting of 10, 200, and 406 articles, referred to as $DICE_{10}$, $DICE_{200}$, and $DICE_{406}$, respectively. These subsets are used for different purposes within the methodology, as described in the following subsections.

Event Schema: We adopt the fine-grained annotation schema from the DICE dataset (Bonisoli et al., 2023), which defines seven core semantic roles for theft events:

- **Author (AUT)**: Individual(s) responsible for the theft.
- **Author Group (AUTG)**: Groups of perpetrators.
- **Victim (VIC)**: Individual(s) affected by the theft.
- **Victim Group (VICG)**: Groups of victims.
- **Location (LOC)**: Geographic or contextual setting where the theft occurred.
- **Object (OBJ)**: Stolen item(s).
- **Injured Party (PAR)**: Commercial establishments or collective entities harmed by the theft.

The schema accounts for complex annotation phenomena including multi-word expressions, multi-span mentions (e.g., coreference), and multi-label assignments (e.g., a location that is also an injured party).

The single-event-per-document assumption reflects a methodological choice that isolates the argument extraction task, abstracting away from event detection and classification. Our approach concentrates on extracting the event-specific arguments, while providing a controlled and high-quality benchmark for the theft event domain.

4.1. Phase 1: LLM evaluation for crime news generation

4.1.1. Input

For the validation phase (Section 3.3.1), we used the news articles from **$DICE_{406}$**. These articles are coupled with their corresponding JSON files and served as the structured input for evaluating candidate Content Generator LLMs. This sample provides a balanced representation of theft scenarios while being of manageable size for thorough validation.

4.1.2. Content generation

We selected seven instruction-tuned Content Generators LLMs for Phase 1. The first model is **Llama-3-8B-Instruct (LL3-8B)**, an improved multilingual model over Llama2, enhanced through high-quality instruction tuning (Grattafiori et al., 2024). We also include **LLaMAntino-3-ANITA-8B (LLT3-8B)**, a variant of LL3-8B fine-tuned specifically on Italian datasets (Polignano et al., 2024). Another selected model is **Mistral-7B-Instruct-v0.3 (M7B)** (Jiang et al., 2023), a compact and efficient 7-billion parameter model. **Zephyr-7B-beta (Z7B)** is also included, built on Mistral-7B and further fine-tuned on a mix of publicly available synthetic datasets. Besides, we consider two instruction-tuned models from Alibaba Cloud's Qwen2.5 series, namely **Qwen2.5-7B-Instruct (Q7B)** and **Qwen2.5-14B-Instruct (Q14B)** (Cloud & Team, 2024), which are based on the LLaMA architecture and leverage sparse attention and extended context length to improve inference efficiency. Finally, we include **DeepSeek-V2 Chat (DS)** (Team et al., 2024), a transformer-based, instruction-tuned large language model employing a Mixture-of-Experts (MoE) architecture, enabling high model capacity with sparse computation. The generation tasks were carried out with constrained computing resources, executing the models on one NVIDIA A100 40 GB GPU, except the Qwen models that were run on one NVIDIA H100 94 GB GPU and DeepSeek which was run with 4-bit quantization on two H100 94 GB GPU.

Generation prompt design (v1.0): The Content Generator LLMs are prompted (prompt v1.0) to generate Italian crime news articles that (i) strictly incorporate all the information contained in a structured JSON, without introducing extraneous content, and (ii) preserve fluency and a journalistic writing style. The prompt first presents the structured JSON, which specifies the annotations that must be included in the generated news article. The JSON consists of seven key–value pairs, corresponding to the seven labels of the DICE schema. The JSON format was selected due to its standardized structure and flexibility, which facilitate the representation of hierarchical data and variable-length annotations. In each key–value pair, the key corresponds to a label name, while the value contains the associated annotations. Values may take different forms depending on the label: a single string when representing one span for a single entity, a list of strings when multiple spans are associated with a single entity, or a list of lists of strings when representing multiple entities under the same label (e.g., multiple victims mentioned in the same crime report). Following the JSON input, the prompt provides explicit generation constraints, e.g., to generate a single theft-related news article written in Italian, to include all strings contained in the JSON, and to avoid adding any information not explicitly specified in the JSON. The prompt also includes a brief explanation of each DICE label, as well as additional clarifications addressing potentially ambiguous distinctions, such as those between `AUT` and `AUTG`, and between `VIC` and `VICG`.

Three prompting strategies are explored: zero-shot, one-shot, and two-shot learning. In the one-shot and two-shot settings, the prompt additionally includes one or two real crime news articles, respectively, selected from $DICE_{10}$. Each example is provided together with its manually annotated JSON representation, serving as guidance for the generation task.

4.1.3. Validation

To evaluate the quality of synthetic news generated by the seven Content Generator LLMs, we employed the multi-dimensional framework described in Section 3.6, categorizing metrics into five functional dimensions: Semantic Alignment, Textual Quality, Lexical Diversity, and Operational Efficiency. The evaluation compares the generated text with the original $DICE_{406}$ dataset to assess realism, while faithfulness is measured by extracting structured information from the generated news using Mixtral-8x7B-Instruct as the Data Extractor LLM (Bonisoli et al., 2025; Jiang et al., 2024). Due to limited computational resources, we ran the model with 4-bit quantization on two NVIDIA A100 40 GB GPUs using the Hugging Face transformers library.[7]

An analysis of the generated outputs focused on language and output format: although Italian free-text responses were expected, minor deviations were observed. Specifically, Z7B produced non-Italian outputs in approximately 2%–3% of cases (some news are in Spanish or in English) and structured JSON outputs in up to 1.5% of documents, while Q7B showed non-Italian generations in less than 1% of cases (some news are in Chinese). Nevertheless, these models were retained as a simple filtering step effectively removes such spurious outputs.

Table 2 summarizes the models' performance across zero-, one-, and two-shot configurations (column #ex). Following the taxonomy in Table 1, the results reveal several key insights regarding the trade-off between efficiency, realism, and structural precision:

Semantic Alignment: All models achieve high sentence-level similarity ($sBS > 60\%$, peaks at 74%), ensuring the core event meaning is preserved. However, the relatively low BERTScore (BS) is an intentional feature of our framework: it reflects the models' ability to generate novel journalistic narratives around fixed facts rather than merely paraphrasing the original text, thus increasing the lexical robustness of the final synthetic dataset.

Textual Quality: The synthetic news generally matches or exceeds the readability of the original DICE dataset. All models achieved Gulpease Index (GI) scores between 51–56 (compared to 50 for DICE), indicating high comprehensibility. Perplexity (PLX)

[7] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

Table 3
Frequency of gender and nationality across randomly generated JSONs.

	AUT	AUTG	VIC	VICG
# entities	2858	6279	4275	447
Male (%)	26	28	31	8
Female (%)	24	12	28	10
Neutral (%)	50	60	41	82
# nationalities	179	970	659	0
Asia (%)	28	25	26	–
Oceania (%)	10	3	5	–
Africa (%)	27	30	27	–
Europe (%)	21	27	30	–
America (%)	14	15	12	–

results show that DS (5.8–6.7) and Qwen models produce the most fluent and predictable text, closely aligning with the human-written benchmark (PLX: 9.5). Conversely, LLT3-8B exhibits higher perplexity (16.2–18.1), suggesting more creative or varied phrasing.

Lexical Diversity: While the original DICE dataset shows near-perfect diversity (*Div-3*: 1.00), synthetic models start slightly lower in zero-shot. However, the introduction of few-shot examples effectively bridges this gap, bringing diversity scores (0.95–0.99) nearly on par with real-world data. This demonstrates that providing examples helps LLMs avoid repetitions and promotes a wider variety of event expressions.

Structural Aligment: This dimension is critical for the usability of the dataset in event extraction. Qwen and DS models demonstrate superior grounding, with *omit* rates often below 30%, leading to the highest EM and PM scores (up to 75.6). While few-shot prompting generally reduces omissions for most models, Qwen and DS perform optimally even in zero-shot, indicating they are better pre-trained for strictly following structured annotation schemas.

Operational Efficiency: Generation time (*gen*) varies significantly: Qwen models are the most efficient, producing long narratives (up to 1140 characters) in less than 2 s. In contrast, DS requires over 75 s per item. These metrics confirm the practical scalability of our pipeline, as even the slowest model is orders of magnitude faster than manual human drafting.

4.1.4. LLMs selection

Considering all metrics, Q7B and Q14B stand out as the best content generators, achieving high JSON faithfulness, low perplexity, good diversity in few-shot, and fast generation. DS also performs well, particularly in semantic alignment. Based on this analysis, we define our model selection for the next phase of synthetic dataset creation. This set includes the top performers Q7B, Q14B, and DS. Furthermore, we also include LL3-8B: while its JSON faithfulness is lower, it exhibits distinct characteristics such as higher text perplexity and shorter generation length. Including a model with a different performance profile will allow for a more comprehensive evaluation of how generator traits influence the final synthetic dataset quality in subsequent stages.

4.2. Phase 2: Synthetic dataset creation, description and evaluation

4.2.1. Entity description generation

To populate the novel SYNTH-ITA corpus of news (Section 3.3.2) with realistic theft scenarios, we first create **curated lists of realistic entities**, where each entry is a plausible textual span corresponding to a specific role label.

These lists are organized in structured files. For the `LOC` role, categories such as parks, shops, and offices are associated with plausible stolen objects and paired with real-world addresses or place names from the province of Modena. For `PAR`, lists include commercial establishments (e.g., bars, companies, retail stores) and their associated plausible stolen objects. `OBJ` spans are grouped according to the type of location or commercial establishment they are typically found in. Attributes for perpetrators (`AUT/AUTG`) and victims (`VIC/VICG`), such as names, nationalities, and places of origin, are also procedurally generated. These comprehensive lists serve as the foundational pool from which entity descriptions are dynamically sampled during the scenario generation phase.

4.2.2. Randomized JSON generation

A dedicated Python library orchestrates the generation of realistic, pseudo-random theft annotations, structuring them as JSON objects. The process sequentially selects: (i) the victim type (individual or organization), (ii) the stolen objects, and (iii) the specific attributes for perpetrators, victims, and locations, drawing from the curated entity lists defined in Section 4.2.1.

To proactively mitigate input-level biases in the synthetic corpus, the generation algorithm incorporates **fairness safeguards** by enforcing balanced distributions for sensitive attributes, specifically the gender and nationality of perpetrators and victims.

4.2.3. Synthetic news generation

We aimed to generate 10,000 synthetic news articles to create a resource that aligns with the scale of related benchmarks (e.g., MultiSpanQA (H. Li et al., 2022), DICE). This volume ensures a sufficient data foundation to be useful for other researchers and to support stable model training for downstream tasks. Moreover, generating 10,000 items provides an order of magnitude more training examples than the available manual annotations for theft. This deliberate scaling allows us to investigate a key research question: whether a larger, automatically generated dataset can lead to improved downstream task performance compared to a smaller set of high-quality manual annotations.

To achieve this, we first generated a corresponding corpus of 10,000 structured JSON annotations (as detailed in Section 4.2.2). To validate the effectiveness of the randomized JSON generation, we conducted a statistical analysis of the generated JSONs. We verified that no specific nationality or gender was systematically overrepresented. Nationalities were grouped into five broad geographic regions, and their frequencies were calculated. Similarly, the distribution of gender assignments (Male, Female, Neutral) was analyzed. The results, presented in Table 3, confirm the balanced nature of the generated attributes across the different semantic roles. Since in Phase 1 performance improves with a larger number of examples, we utilized LL3-8B, Q7B, Q14B and DS with 4-shot learning with examples derived from $DICE_{10}$. Using the 10,000 JSONs as input and a new generation prompt (v1.1), each LLM generated 10,000 documents (mirroring Phase 1's methodology).

Generation prompt design (v1.1): Prompt v1.1 is defined based on the evaluation of the news articles generated in Phase 1 using prompt v1 described in Section 4.1.2. A more concise and structured formulation is introduced, explicitly enforcing verbatim inclusion of all JSON spans, clarifying the handling of empty fields, and formalizing logical dependencies among labels (e.g., if `PAR` is present, `VIC` is empty in the JSON and cannot be mentioned in the generated text). These refinements aim to reduce unintended paraphrasing, improve text quality and enhance JSON faithfulness. This process resulted in the creation and public release of four distinct SYNTH-ITA v1.1 (SYN1) datasets, each generated by one of the selected LLMs.

4.2.4. Iterative manual evaluation and prompt refinement

To assess and iteratively improve output quality, we conducted a manual evaluation of the generated news. Two domain experts independently analyzed and annotated 80 synthetic news generated with prompt v1.1: 40 with complete JSON information and 40 with omitted spans. The inter-annotator agreement, measured by Krippendorff's α (Krippendorff, 2006), was 0.826, indicating substantial reliability.

This review served to verify real-world plausibility, identify potential biases, and check adherence to annotation guidelines. The manual annotations showed strong alignment with the original JSONs, with EM scores of 89.4% and PM scores of 92.6%, respectively, confirming high structural fidelity.

The evaluation revealed recurring issues, which we categorized as linguistic or JSON-related. Linguistic issues included:

- **Grammatical and syntactic errors**: the generated texts frequently exhibit mistakes in grammar and syntax, e.g., improper use of articles.
- **Inconsistent or incoherent information**: contradictory or unclear information, such as: *"The perpetrators of the theft have been identified as a man from the Aosta Valley, residing in Nonantola, without further details on his identity".* creates confusion regarding the individuals involved.
- **Redundant and unusual expressions**: the generated content often includes repetitive expressions, such as *"The identity of the theft perpetrators is still unknown, as they have not been identified".* or unusual expressions, such as: *"The commercial entity or public institution involved in the theft has not been specified"*..
- **Lack of context**: the generated reports often lack contextual details about the crime's progression, leading to simplistic narratives that deviate from the realistic tone proper of crime reporting.

JSON-related issues included:

- **Label misinterpretation and contextual inconsistency**: the model often misinterprets missing or present fields in the JSON input, leading to semantically inconsistent or misleading statements. This includes confusing individual and group entities for `AUT` or `VIC`, resulting in incoherent and contradictory text. For example, when `AUT` is specified and `AUTG` is empty in the JSON, the generated news correctly describes the perpetrator but an unusual expression like *"the group of thieves is unknown"* is present.
- **OBJ description**: the term *"precious objects"* is frequently overused without clear contextual relevance. Additionally, item lists often contain redundancies or semantically incongruent combinations, such as *"two mobile phones, a mobile phone"* in the same news.
- **LOC ambiguity**: the representation of crime locations is often vague or incomplete. For example, if a theft is reported as having occurred *"on a street"*, it remains unclear whether the incident occurred on the public highway, inside a building adjacent to the street or in another context.

These findings underscored the need to refine the generation prompt (leading to the development of a new version, v2.0), strengthen the model's capacity to accurately handle and integrate structured data, and enhance the naturalness and coherence of generated texts.

Generation prompt design (v2.0): With respect to v1.1, the new prompt prioritizes journalistic realism by encouraging the model to add contextual framing and a description of the event dynamics and respect the event-level coherence around the theft. Prompt v2.0 presents a simplified set of instructions that omits several low-level constraints, such as the explicit dependency between `PAR` and `VIC` and the exhaustive enumeration of socio-demographic attributes.

4.2.5. Final dataset generation

Using prompt v2.0, we executed the final synthetic news generation. This process followed the same methodology detailed in Section 4.2.3: each of the four selected LLMs (LL3-8B, Q7B, Q14B, and DS) was used to generate 10,000 news articles, utilizing the same 10,000 JSON as input but with the improved prompt v2.0. This step resulted in the creation of four new, publicly released SYNTH-ITA v2 (SYN2) datasets.

Table 4
Analysis of publishable news instances ($\#ne_{no\text{-}omit}$) in the SYNTH-ITA v2 (SYN2) datasets, before (BP) and after (AP) post-processing.

	Metric	SYN2			
		LL3-8B	Q7B	Q14B	DS
BP	$\#ne_{generated}$	10,000	10,000	10,000	10,000
	$\#ne_{format\ issue}$	0	2288	0	2
	$\#ne_{non\text{-}Italian\ text}$	0	91	0	0
	$\#ne_{1+omitted\text{-}spans}$	6582	5360	7665	6869
	$\#ne_{no\text{-}omit}$	3418	2261	2335	3129
AP	$\#ne_{no\text{-}omit} = \#ne_{released}$	**7534**	**4957**	**6238**	**6990**
	Acceptance rate ($\#ne_{released}/\#ne_{generated}$)	75%	50%	62%	70%

Table 5
Evaluation metrics ($\#ne_{ann}$, EM (%), PM (%)) for synthetic datasets SYN2.

Metric	SYN2			
	LL3-8B	Q7B	Q14B	DS
$\#ne_{ann}$	80	80	80	80
EM (%)	83.2	84.1	85.2	79.0
PM (%)	88.2	86.9	89.9	89.8

4.2.6. Post-processing

A dedicated post-processing step was applied to each of the four SYNTH-ITA v2 datasets to ensure high fidelity between the generated text and the original JSON annotation. This verification and correction process is critical for creating a reliable, aligned resource for downstream information extraction tasks.

Prior to the main alignment check, articles with severe generation errors were filtered out. As shown in Table 4, this included items with format issues (e.g., JSON outputs instead of text) or non-Italian text, which are unusable for the intended task. This initial filtering explains the reduction from the original 10,000 generated items to the total generated counts considered for alignment. For instance, Q7B showed notable instability, with a significant portion of outputs containing format errors or non-Italian text, which were consequently removed.

The core post-processing, detailed in Section 3.4, involves a rule-based, span-by-span alignment verification. For each news item, the algorithm: (1) checks for the presence of each JSON span in the text, (2) searches for semantically equivalent formulations (e.g., synonyms, morphological variants) if a literal match is absent, (3) updates the JSON span to the exact text mention when a match is found. Spans for which no textual equivalent can be identified are flagged as omitted.

A conservative filtering policy was then applied to these omitted spans to preserve as much data as possible while ensuring event integrity and realistic reporting scenarios. Omissions of non-critical details (e.g., a description of a perpetrator within a `AUT` span) led only to the removal of that span from the JSON. However, if a core event argument was missing (e.g., the location of the event `LOC`), the entire news item was discarded. Moreover, the absence of a stolen object (`OBJ`) is not automatically considered a critical omission if the location (`LOC`) is a private, enclosed space (e.g., a house or apartment), reflecting real cases where a burglary is reported before a full inventory is possible.

This process created the final high-fidelity subset for each dataset, ensuring that plausible theft narratives are retained while **unrelated events**, which lack the fundamental arguments of the theft schema, are systematically filtered out.

The effectiveness of this post-processing is evident in the final rows of Table 4 ($\#ne_{no\text{-}omit}$), which show the number of news articles with perfect alignment before and after processing. While a majority of initial articles contained omitted spans (65%–77%), post-processing successfully corrected or filtered these cases, significantly increasing the yield of perfectly aligned news. The acceptance rate, defined as the fraction of generated articles that survive post-processing, varies substantially across models (50%–75%). This variability in acceptance rate should be interpreted alongside the generation times (reported in Table 2): while Q7B exhibited the fastest generation (1.8–5.9 s per article), its high filtering loss reduces the effective throughput, whereas LL3-8B, despite being slower, yields a larger final corpus.

4.2.7. Evaluation

Expert Evaluation: Structural Alignment and Narrative Coherence

A manual evaluation was conducted to assess the realism and structural fidelity of the post-processed SYN2 datasets. To enable a direct analysis of narrative improvement prompted by the same structured input, we generated the new sample using the same 80 JSONs from the previous evaluation in Section 4.2.4, this time with the final prompt v2.0. Domain experts independently annotated this new set of 80 news ($\#ne_{ann}$) from each SYN2 dataset, and these manual annotations were compared against the post-processed JSONs.

The results, presented in Table 5, show high EM and PM scores (consistently > 83%), confirming a good alignment between the generated narratives and the intended event structures. The scores are slightly lower than those reported in Section 4.2.4, which is expected: the previous evaluation used a balanced sample (40 news with and 40 without omitted spans), while this final sample

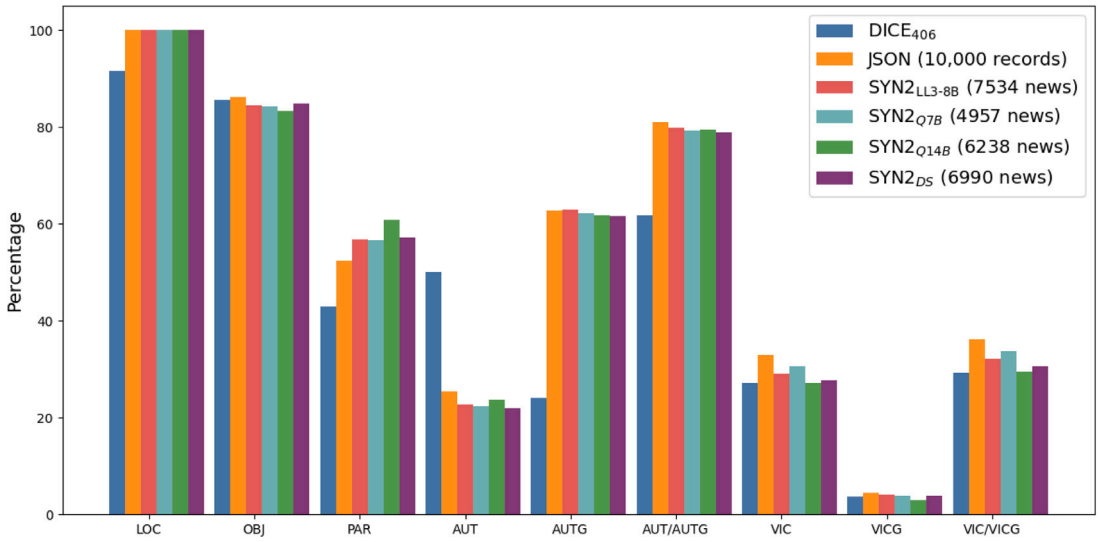

Fig. 2. Distribution of label occurrence across real (DICE), randomly generated JSON, and synthetic news (SYN2) datasets.

was naturally unbalanced, containing a higher proportion of articles that originally had omitted spans, a more challenging set for alignment.

Critically, experts noted substantial improvements in narrative coherence and detail over the v1.1 datasets, validating the prompt refinement. Compared to the previous version, enhancements included clearer context introduction, more detailed descriptions of theft dynamics, and the addition of narrative elements (e.g., damaged objects, getaway vehicles, or secondary characters like witnesses) not explicitly provided in the JSON input.

Despite these improvements, certain issues persisted from prompt v1.1, including repetitions, inconsistencies, grammatical errors, and confusion between semantic roles such as AUT and VIC. In few cases, the increased narrative flexibility in v2.0 introduced new types of errors. Approximately 3% of stories contained implausible event sequences (e.g., "calmly inviting a suspicious stranger inside"). Model-specific failures also emerged, such as LL3-8B generating completely unrelated events (e.g., accidents or assaults) in about 10% of its outputs.

Label Distribution Fig. 2 compares the percentage of news articles containing each label across the original DICE_{406} dataset, the randomly generated JSON records, and the different synthetic datasets.

Post-processing increases the number of news articles without omitted spans, at the cost of reducing the overall label coverage. LOC is consistently the label with the lowest omission rate across all datasets. When omissions occur in synthetic news, they typically involve a single span: 3% of spans are missing for $\text{SYN2}_{\text{LL3-8B}}$, compared to 12% for Q7B and 6% for Q14B.

Overall, the label distributions observed in the JSON records are largely preserved in the synthetic datasets. This preservation is achieved by enforcing distributional constraints at the level of semantically related role groups, namely AUT/AUTG and VIC/VICG, rather than by independently controlling the frequency of each fine-grained label. This indicates that, after the JSON update performed during post-processing, even when some spans are removed, at least one span per label is retained. Moreover, no substantial differences in label distributions are observed among the four synthetic datasets.

Across all datasets, LOC and OBJ emerge as the most frequent labels. In particular, both labels are present in nearly all JSON records and are consistently preserved in the synthetic datasets, closely matching the distribution observed in real news articles. This reflects the structural properties of theft-related reporting, which almost invariably includes information about the location of the crime and the stolen objects.

AUTG and PAR are more prevalent in the synthetic datasets than in DICE_{406}, mirroring their higher frequency in the JSON inputs.

The largest discrepancies between real and synthetic data concern the AUT label, a consequence of the grouped distributional constraints adopted during JSON generation. In DICE_{406}, AUT appears in about half of the articles, whereas it is present in roughly one quarter of the JSON records and consequently in a similarly reduced proportion of synthetic news. Considering the percentage of news articles in which at least one label between AUT and AUTG appears (the AUT/AUTG group in the diagram), the proportion across real and synthetic datasets is more stable.

VIC and VICG show relatively stable proportions across real, JSON, and synthetic data. VICG remains rare across all datasets, confirming that group-level victim mentions are uncommon in theft-related news. If we consider the VIC/VICG group in the diagram, we observe that this group remains the least frequent.

Table 6 presents a comprehensive comparison between the real DICE dataset and the four synthetic SYN2 datasets across multiple evaluation dimensions defined in Section 3.6.

The synthetic datasets provide a substantial increase in scale, with each SYN2 yielding between 4957 and 7534 high-fidelity news items, an order of magnitude larger than the manually annotated theft subset in DICE (606 items).

Table 6
Comparison between real (DICE) and synthetic (SYN2) datasets released.

Dataset	Textual Quality						Diversity										
	$\#ne_{released}$	len	len_{sen}	voc	GI	PLX	Dist-2	Dist-3	Div-2	Div-3	JSD-2	JSD-3	MTLD	HD-D	MATTR	Self-BLEU	Self-Repetition
$DICE_{10,395}$	**10,395**	1874	28	192	50	10.1	0.96	0.99	0.97	1.00	–	–	115	0.85	0.84	0.02	1.00
$DICE_{406}$	**406**	1489	23	160	50	9.5	0.96	0.99	0.97	1.00	0.45	0.75	113	0.85	0.84	0.02	0.99
$SYN2_{LL3\text{-}8B}$	**7534**	1059	23	113	52	7.2	0.91	0.97	0.88	0.96	0.62	0.85	75	0.80	0.79	0.02	1.00
$SYN2_{Q7B}$	**4957**	1138	25	120	46	9.4	0.94	0.98	0.90	0.97	0.63	0.86	88	0.82	0.80	0.01	1.00
$SYN2_{Q14B}$	**6238**	907	23	106	46	9.2	0.97	0.99	0.90	0.97	0.63	0.86	97	0.83	0.82	0.01	1.00
$SYN2_{DS}$	**6990**	1219	26	123	49	7.2	0.93	0.98	0.89	0.96	0.64	0.85	84	0.82	0.81	0.01	1.0

Textual Quality The synthetic news are consistently shorter than real DICE articles (907-1219 vs. 1874 characters), yet they maintain comparable sentence structure (23–26 words per sentence) and high readability (Gulpease Index > 46). Perplexity scores reveal a trade-off: models with lower perplexity (LL3-8B and DS at 7.2) produce more fluent but potentially less varied text, whereas those with perplexity closer to the real data (Q7B and Q14B at 9.2–9.4) better mimic the linguistic unpredictability of human-written news. An additional Part-of-speech (POS) analysis further confirms the grammatical soundness of the synthetic texts. The frequencies of verbs and nouns remain stable and closely aligned with the real DICE subset (across SYN2: verbs 8%–11%, nouns 22%–24%; $DICE_{406}$: verbs 11.22%, nouns 21.26%). While adjectives show moderate model-dependent variability (e.g., 6–7.6% for Q14B and DS vs. 4.6% in DICE), core grammatical elements such as adpositions and determiners are stable indicating that the synthetic texts successfully replicate the fundamental grammatical framework of real news.

Lexical Diversity We evaluate lexical diversity at three levels: within individual documents, across the dataset, and between datasets. *Within-document* diversity is high across all datasets: Dist-2 and Dist-3 consistently exceed 91% and 97%, indicating that synthetic articles avoid local repetition. However, lower MTLD values (75–97) compared to real DICE (113) reveal a limitation in sustaining lexical variety over longer narrative spans. *Across the dataset*, Div-N scores approach near-perfect novelty (Div-2: .88-.90, Div-3: .96-.97), confirming that articles are phrased distinctly, mitigating the risk of repetitive patterns or data leakage. *Cross-dataset* divergence, measured by JSD-N, shows a measurable shift from real news. Notably, divergence increases with n-gram length, indicating that differences in syntactic structures and narrative flow are more pronounced than in local word co-occurrence. Finally, *repetition* metrics (Self-BLEU and Self-Repetition) align with journalistic conventions: low Self-BLEU reflects inter-document distinctiveness, while high Self-Repetition captures the natural reuse of common phrases (e.g., “according to police reports”, “the investigation is ongoing”.).

4.2.8. Released

The four synthetic datasets, validated and cleaned through the proposed post-processing pipeline, together with the source code of our methodology and the generation prompts, are released to support the training and evaluation of document-level event extraction models and are available at https://github.com/federicarollo/SYNTH-ITA.

5. Downstream evaluation: Event argument extraction with QA and NER formulations

SYNTH-ITA can be used as training data for pre-training, fine-tuning, and evaluation in various Italian NLP tasks. In this work, we specifically assess the impact of SYN2 on event argument extraction for Italian theft events, employing two alternative modeling formulations: Question Answering-formulated Event Extraction (QA-EE) and Named Entity Recognition-formulated Event Extraction (NER-EE). These are not general-purpose QA or NER tasks, but rather specialized adaptations designed to extract the seven semantic roles defined in our theft event schema.

For QA-EE, we create natural language questions for each event role (e.g., “What was stolen?” for OBJ). For NER-EE, we treat each semantic role as a distinct entity type. This framework allows us to compare how different modeling paradigms perform on the same underlying event extraction problem.

Specifically, we conducted two complementary experiments. The first experiment (Test 1) aims to compare how QA-EE and NER-EE models trained on synthetic data perform relative to those trained on general-domain or smaller manually annotated real datasets when evaluated on a realistic target dataset. In this experiment the test set is $DICE_{406}$. The second experiment (Test 2) isolates the effects of domain alignment by testing the models on the manually annotated subset of synthetic news, which acts as a gold standard. The evaluation has been conducted considering EM and PM as defined in Section 3.6. Together, these tests provide a foundation for evaluating the trade-offs between scalability and linguistic authenticity in event extraction systems.

Table 7
TEST 1 - Event argument extraction performance on real $DICE_{406}$ news: QA-EE vs NER-EE models trained on various datasets. Metrics include precision (P), recall (R) and F1 score. The best scores for each model type are shown in bold, while the second-best scores are underlined.

	Training dataset	EM (%)			PM (%)		
		P	R	F1	P	R	F1
BERT	SimplePrompt	41.1	33.9	37.2	51.2	41.1	45.6
	$MSQA_{ita}$	35.3	23.6	28.3	38.5	26.0	31.0
	$DICE_{10}$	–	–	–	–	–	–
	$DICE_{80}$	35.7	22.9	28.0	38.1	24.2	29.6
	$DICE_{200}$	**46.8**	45.0	**45.0**	**59.0**	49.3	53.8
	$SYN2_{manual\text{-}80}$	32.3	32.5	32.4	48.6	39.7	43.7
	$SYN2_{LL3\text{-}8B}$	44.5	**43.5**	44.0	56.7	51.0	53.7
	$SYN2_{Q7B}$	44.2	**43.5**	43.8	57.8	51.7	**54.6**
	$SYN2_{Q14B}$	41.6	42.0	41.8	55.7	50.3	52.9
	$SYN2_{DS}$	45.1	42.5	43.8	49.6	**56.2**	52.7
ELECTRA	SimplePrompt	40.4	37.2	37.7	50.3	44.2	47.0
	$MSQA_{ita}$	35.6	24.1	28.7	42.1	30.1	35.1
	$DICE_{10}$	–	–	–	–	–	–
	$DICE_{80}$	17.7	24.0	20.4	24.0	24.4	24.1
	$DICE_{200}$	43.9	36.2	39.7	57.4	41.6	48.3
	$SYN2_{manual\text{-}80}$	31.3	24.0	27.2	33.6	24.1	28.1
	$SYN2_{LL3\text{-}8B}$	46.5	**44.5**	**45.5**	58.0	51.8	**54.7**
	$SYN2_{Q7B}$	44.7	43.6	44.1	57.5	51.9	54.5
	$SYN2_{Q14B}$	42.9	43.3	42.9	57.2	**52.2**	54.6
	$SYN2_{DS}$	**46.9**	43.2	45.0	**59.1**	50.6	54.5
SpacyNER	SimplePrompt	51.4	45.3	45.6	56.7	50.2	53.3
	$DICE_{10}$	53.3	44.6	48.6	57.5	46.5	51.4
	$DICE_{80}$	58.2	55.7	57.0	64.0	60.5	62.2
	$DICE_{200}$	**59.6**	**61.0**	**60.3**	**66.8**	**65.8**	**66.3**
	$SYN2_{manual\text{-}80}$	42.8	49.7	46.0	50.9	38.9	45.1
	$SYN2_{LL3\text{-}8B}$	53.2	47.1	50.0	58.4	52.1	55.1
	$SYN2_{Q7B}$	52.8	42.9	47.4	56.8	46.9	51.4
	$SYN2_{Q14B}$	49.8	50.8	50.3	57.5	56.0	56.7
	$SYN2_{DS}$	51.1	48.3	49.7	57.9	53.6	55.7

5.1. Selected models

We employed multi-span QA models based on Italian instances of BERT[8] (Devlin et al., 2019) and ELECTRA[9] (Clark et al., 2020), as well as the SpaCyNER model from the Italian `it_core_news_md` pipeline.[10] The SpacyNER model requires fine-tuning on annotated data aligned with the target entity types.

To adapt the datasets for QA-EE models, a natural language question was formulated for each label (e.g., "What was stolen?" for `OBJ`), and paired with the corresponding news article and annotated answer. In addition, for QA-EE models, the `AUT` and `AUTG` labels were grouped and evaluated using a single question ("Who is the thief or the criminal?"), since the answers can refer to individuals or groups. The same approach was adopted for `VIC` and `VICG`. In contrast, the NER-EE model evaluated `AUT`, `AUTG`, `VIC`, and `VICG` separately. However, for easier comparison across models, performance metrics are reported for the label pairs (`AUT/AUTG`) and (`VIC/VICG`) in the following sections.

For both QA-EE and NER-EE, we fine-tuned the models using varying amounts of real annotated data from DICE ($DICE_{10}$, $DICE_{80}$[11] and $DICE_{200}$), as well as the manually annotated 80-synthetic-news portion ($SYN2_{manual\text{-}80}$) and compared these against models fine-tuned on different variants of SYN2 ($SYN2_{LL3\text{-}8B}$, $SYN2_{Q7B}$, $SYN2_{Q14B}$ and $SYN2_{DS}$). For QA-EE, we also fine-tuned the models using $MSQA_{ita}$, a multi-span generic-domain dataset consisting of 10,295 question-answer pairs, obtained via automatic translation of the English dataset MultiSpanQA (H. Li et al., 2022).

[8] https://huggingface.co/mrm8488/bert-italian-finedtuned-squadv1-it-alfa
[9] https://huggingface.co/anakin87/electra-italian-xxl-cased-squad-it
[10] https://spacy.io/models/it
[11] $DICE_{80}$ comprises 80 news articles selected from $DICE_{200}$. This subset is included to enable a direct comparison with $SYN2_{manual\text{-}80}$, which contains the same number of news articles.

Table 8
TEST 1 - Label-wise F1 scores on PM for QA-EE and NER-EE formulations on $DICE_{406}$ (numbers inside brackets exclude news with empty annotations for that label from the ground-truth). The best results per label-model pair are shown in bold, while the second-best scores are underlined.

		LOC	OBJ	PAR	AUT/AUTG	VIC/VICG
				Label statistics		
	#annotated spans	686	664	175	675	203
	% news with annotations	92	85	43	74	31
	Training dataset			**F1-score (%)**		
BERT	$MSQA_{ita}$	15.8 (11.3)	19.4 (10.4)	56.6 (0.1)	23.8 (0.1)	51.7 (14.6)
	$DICE_{10}$	–	–	–	–	–
	$DICE_{80}$	11.6 (6.0)	10.5 (0.6)	54.7 (0.8)	24.7 (1.3)	60.3 (1.1)
	$DICE_{200}$	**65.5** (**67.5**)	51.3 (49.9)	67.3 (46.3)	25.1 (0.1)	**64.1** (1.1)
	$SYN2_{manual\text{-}80}$	45.2 (46.2)	47.3 (47.8)	57.8 (1.2)	20.9 (4.8)	55.2 (7.3)
	$SYN2_{LL3\text{-}8B}$	62.9 (65.4)	47.2 (45.3)	67.8 (43.0)	37.4 (29.9)	58.5 (42.8)
	$SYN2_{Q7B}$	61.4 (63.8)	50.4 (45.3)	67.5 (39.6)	**46.5** (**43.1**)	50.1 (**49.9**)
	$SYN2_{Q14B}$	61.5 (63.9)	**51.9** (**52.8**)	66.4 (40.2)	33.0 (25.7)	56.7 (49.0)
	$SYN2_{DS}$	64.3 (66.7)	49.2 (46.2)	**70.8** (61.0)	37.9 (27.4)	56.7 (45.8)
ELECTRA	$MSQA_{ita}$	30.1 (26.0)	27.7 (20.7)	53.9 (11.1)	23.0 (8.4)	51.3 (10.6)
	$DICE_{10}$	–	–	–	–	–
	$DICE_{80}$	7.4 (1.7)	11.4 (1.4)	39.4 (0.5)	20.9 (2.0)	46.1 (3.0)
	$DICE_{200}$	**66.4** (67.6)	25.2 (18.1)	67.7 (31.3)	24.1 (0.4)	–
	$SYN2_{manual\text{-}80}$	15.7 (11.3)	33.0 (30.8)	57.1 (10.8)	22.1 (7.9)	48.4 (4.4)
	$SYN2_{LL3\text{-}8B}$	64.0 (66.6)	46.9 (44.8)	69.0 (39.5)	**41.0** (36.2)	**58.0** (51.0)
	$SYN2_{Q7B}$	62.7 (65.4)	52.5 (51.6)	**70.5** (49.9)	40.4 (36.2)	51.0 (52.1)
	$SYN2_{Q14B}$	62.4 (65.0)	**52.6** (54.1)	69.4 (47.1)	38.0 (7.9)	55.7 (45.6)
	$SYN2_{DS}$	62.8 (64.4)	49.2 (48.6)	68.6 (38.1)	39.0 (33.7)	50.9 (47.1)
SpacyNER	$DICE_{10}$	46.1 (47.0)	–	–	–	–
	$DICE_{80}$	55.2 (56.5)	30.4 (24.8)	56.5 (0.3)	72.5 (67.77)	79.3 (10.6)
	$DICE_{200}$	52.6 (52.3)	**49.8** (47.7)	61.6 (16.3)	**74.2** (72.2)	**81.0** (11.0)
	$SYN2_{manual\text{-}80}$	50.8 (51.9)	41.9 (38.7)	60.0 (18.1)	72.8 (70.1)	79.8 (45.6)
	$SYN2_{LL3\text{-}8B}$	50.7 (50.9)	37.8 (34.4)	61.5 (19.1)	59.6 (26.2)	61.7 (41.7)
	$SYN2_{Q7B}$	33.7 (32.1)	36.3 (33.0)	57.1 (19.1)	56.8 (26.2)	64.7 (28.0)
	$SYN2_{Q14B}$	**58.8** (61.3)	46.8 (46.6)	–	57.9 (30.8)	61.1 (42.8)
	$SYN2_{DS}$	47.3 (47.9)	43.0 (43.6)	**62.4** (28.0)	60.8 (37.9)	61.6 (42.2)

5.2. Test 1 - generalization on real data

Test 1 evaluates the real-world applicability of SYN2 by measuring how well models trained on synthetic data generalize to real, manually annotated crime reports. Table 7 summarizes the overall performance of the QA-EE and NER-EE models on the $DICE_{406}$ test set (not included in the fine-tuning datasets). A dash indicates that the model failed to correctly predict any annotation for all labels in the documents where they actually occurred.

For QA-EE models, training on any variants of large-scale synthetic data consistently yields strong performance that is competitive with, and in some cases surpasses, training on manually annotated real data. An exception is BERT fine-tuned on $DICE_{200}$, which outperforms models fine-tuned on synthetic datasets in certain cases. However, this advantage comes at the cost of manually annotating 200 news articles, a process that is both time-consuming and resource-demanding. Among the SYN2 variants, performance remains consistent, indicating robustness to the underlying generative model. The QA-EE model trained on the extremely small dataset $DICE_{10}$ fails to generalize and produces only empty outputs. The alignment between the synthetic training data and the DICE test schema is a contributing factor to the strong performance observed. However, this is not a trivial consequence of alignment alone, since neither a small set of perfectly aligned real examples ($DICE_{80}$) nor a large but out-of-domain QA dataset ($MSQA_{ita}$) achieve comparable results.

For NER-EE models, the pattern is different. Here, training on the real $DICE_{200}$ dataset yields the best performance, outperforming all synthetic alternatives by a notable margin (about 10 EM F1 points). This suggests that the sequence-labeling architecture of SpaCy is more sensitive to the distributional shift between synthetic and real text, or benefits more from the precise lexical and syntactic patterns of human writing. Nonetheless, SYN2 datasets still enable the NER-EE model to learn a substantial part of the task, providing a viable baseline when no manual annotations are available.

5.2.1. Naive baseline comparison with a simple-prompt synthetic dataset

To further contextualize the performance of our controlled generation pipeline, we constructed a naive synthetic dataset using a minimal, non-iterative prompt. This baseline serves to quantify the benefits of our structured, document-level, and iteratively refined methodology over a simplistic LLM-based generation approach that does not incorporate explicit alignment mechanisms or expert-driven prompt refinement.

Baseline Generation Methodology We generated a synthetic dataset of 4000 theft-related news articles by providing a straightforward, static prompt to LL3-8B model. The prompt instructed the model to act as a journalist and write an Italian news story based solely on the information contained in a provided JSON structure, without additional constraints or examples. For each of 4000 randomly generated JSONs (we sample these file from the 10,000 JSONs used to generate SYNTH-ITA). A very simple post-processing step filtered out items where any JSON span was omitted from the generated text, ensuring the dataset could be used for training. This process resulted in a final corpus of 4000 aligned (document, JSON) pairs, which we refer to as the *SimplePrompt* dataset.

Quality and Characteristics of the SimplePrompt Dataset Automatic metrics reveal significant qualitative limitations compared to SYN2 datasets (provided in Table 6). The *SimplePrompt* articles exhibit lower lexical diversity (Div-2: 0.83, Div-3: 0.92 vs. SYN2's Div-2: 0.88–0.97, Div-3: 0.96–0.99) and a more constrained vocabulary richness (MTLD: 65.5 vs. SYN2's 75–97.2). The Gulpease readability index (49.8) is comparable, but the narrative structure is markedly inferior.

A **Manual Evaluation** of 24 randomly sampled articles by domain experts identified critical flaws:

- **Rigid and Repetitive Structure:** Articles follow a near-identical template: a bold headline stating location and theft type, followed by date/location, a list of stolen objects, perpetrator/victim details, a standard mention of an ongoing police investigation, and a closing public appeal. This results in extremely low narrative variability.
- **Senseless Sentences and Logical Inconsistencies:** The model often produces contradictory information (e.g., referring to the same perpetrators with conflicting origins) or chaotic, unrealistic lists of stolen items, indicating poor integration of structured data into a coherent narrative.
- **Excessive Lexical Repetition:** Formulaic phrases like "The police are still investigating" or "asks citizens to remain vigilant" appear in almost every article, leading to high self-repetition.
- **Lack of Narrative Fluency:** The output resembles a disjointed list of facts extracted from the JSON rather than a fluid journalistic narrative. Contextual details, logical event progression, and realistic reporting style are largely absent.

Downstream Performance Comparison We fine-tuned the same QA-EE (BERT, ELECTRA) and NER-EE (SpaCy) models on this *SimplePrompt* dataset and evaluated them on the DICE_{406} test set. The results, integrated into Table 7, demonstrate a substantial performance gap.

For QA-EE models, training on the *SimplePrompt* dataset yields significantly lower F1 scores (BERT: EM F1 $\approx$ 37.2, PM F1 $\approx$ 45.6; ELECTRA: EM F1 $\approx$ 37.7, PM F1 $\approx$ 47.0) compared to training on any SYN2 variant (e.g., SYN2_{Q7B} with BERT: EM F1 43.8, PM F1 54.6). The NER-EE model shows a similar trend, with performance closer to that achieved with very small real datasets (DICE_{10}) and far below that of models trained on SYN2.

This comparison underscores that a naive, single-prompt generation strategy produces data of insufficient quality and diversity for effective downstream tasks.

5.2.2. Label-wise performance analysis

To better understand how performance vary on the different labels, Table 8 details label-specific F1 scores for partial match. Each cell reports two values: the overall F1 (calculated over all 406 test documents) and, in parentheses, the F1 computed only on the subset of documents that actually contain at least one ground-truth annotation for that label. This distinction is crucial, as the overall score can be heavily influenced by a model's ability to correctly predict the absence of a label, especially for rare arguments. The header rows provide label statistics: the total number of ground-truth spans and the percentage of documents containing each label.

The performance of the models differs considerably between the labels. For **high-frequency labels** like `LOC` (present in 92% of documents) and `OBJ` (85%), models trained on SYN2 datasets achieve strong overall F1 scores (e.g., 61%–64% for `LOC` with BERT/ELECTRA) that are competitive with models trained on 200 real examples. Among these models, the QA-EE models outperformed NER-EE ones. The scores in parentheses remain similarly high, indicating these models are proficient at both detecting the presence of these common arguments and extracting their correct mentions.

NER-EE models trained on a good amount of real data (DICE_{200}) excel on low-frequency or complex entities such as `AUT/AUTG` and `VIC/VICG`, where synthetic data fall short. This suggests limitations in the coverage of synthetic data and contextual diversity for these labels.

Performance on **rarer labels** presents a more nuanced picture. For `VIC/VICG` (present in only 31% of documents), the overall F1 can be misleadingly high due to correct predictions of label absence. A striking example is the SpaCyNER model trained on DICE_{200}, which achieves an overall F1 of 81.0% but only 11.0% on documents where a victim is actually mentioned. This large gap indicates that the model has learned a conservative strategy, correctly predicting "no victim" in most of the cases, but fails to extract the victim span when it is present. Additionally, QA-EE models trained on small real datasets (DICE_{10} or DICE_{80}) struggle to learn the extraction task for arguments that are actually present and tend to default to predicting label absence as a safe fallback. A similar pattern is observed for NER-EE models trained on DICE_{10}.

For moderately frequent labels, like `PAR` (43% of documents), models trained on SYN2 datasets show a more balanced performance. For instance, BERT trained on SYN2_{DS} achieves an overall F1 of 70.8% and a non-empty F1 of 61.0%, suggesting a reasonable ability to identify when a `PAR` is present and to extract it. In contrast, models trained on out-of-domain data (MSQA_{ita}) or very small in-domain sets (DICE_{10}, DICE_{80}) exhibit drastic drops in the non-empty F1 (e.g., 0.1% for MSQA_{ita} on `PAR`), revealing that they essentially never predict the label when it should be predicted, likely as a safe heuristic to minimize false positives.

The QA-EE models (BERT and ELECTRA) when trained on synthetic data generally show smaller gaps between overall and non-empty F1 for most labels compared to the NER-EE model (SpaCyNER). This suggests that the question-answering formulation provides a more robust mechanism for deciding whether an argument is present or not, rather than defaulting to its absence. For AUT/AUTG, SYN2-trained QA-EE models achieve non-empty F1 scores in the range of 25%–43%, indicating a genuine, albeit modest, extraction capability. In contrast, the NER-EE model trained on real data ($DICE_{200}$) achieves a high non-empty F1 of 72.2% for AUT/AUTG (while for the model trained on $SYN2_{DS}$ is 37.9%), highlighting that sequence labeling can excel when trained on sufficient, high-quality real examples but struggles to transfer from synthetic distributions. Additionally, all QA-EE models when trained on small real datasets ($DICE_{10}$ or ($DICE_{80}$ face big gaps between overall and non-empty F1 across all labels, this is true also for NE-ER on $DICE_{10}$

The label-wise analysis underscores that synthetic data is most effective for learning to extract frequent, central event arguments like location and object. For rarer labels, models, particularly the NER-EE model, tend to learn a conservative strategy that capitalizes on label absence, as reflected in the large disparity between overall and non-empty F1 scores. The QA-EE formulation, although reporting lower F1 score values, demonstrates a advantage in mitigating this issue, yielding more balanced performance across label categories when trained on large-scale synthetic data.

5.2.3. Error analysis of synthetic data trained models

To better understand the nature of failures in models trained on SYN2 datasets, we conducted a manual error analysis. Specifically, given that the SpacyNER model exhibited the greatest performance discrepancy between training on real news and training on the various SYN2 variants, we examined 40 news: 10 news articles from each of the four SYN2 datasets (LL3-8B, Q7B, Q14B, DS) where the model made at least errors in predicting the perpetrator (AUT) label. By examining these cases, we aimed to identify recurring failure patterns and gain insight into the limitations of synthetic data for training document-level event extraction models.

Our analysis revealed that for all models, the main errors fall into three categories:

1. **Perpetrator-Victim Role Confusion (Most Frequent)**: All models systematically mislabel perpetrator attributes as victim descriptors, even when clear contextual cues indicate the perpetrator role. For example, in the text *"The police arrested a 25-year-old Tunisian citizen for attempted theft"*, despite the perpetrator indicators ("arrested for attempted theft"), the demographic attributes are misassigned to the victim (VIC) label.
2. **Group vs. Individual Confusion**: Models frequently mislabel individual descriptors as groups (AUTG/VICG) or viceversa. For instance, an age range descriptor like "between 30 and 40 years old" referring to a single individual is incorrectly extracted as a group label.
3. **Extraneous Object Extraction**: Models often extract containers, tools, or incidental items instead of the actual stolen objects. For example, given the text *"The money had been divided into various hiding places (between a jar of mayonnaise in the refrigerator, a box of hair removal strips and a kitchen cabinet)"*, models tend to extract the hiding places ("refrigerator", "kitchen cabinet") rather than the stolen money.

We observed differences in model behavior: DS and Q14B models tend to extract more entities overall (even if incorrect), while Q7B and LL3-8B are more conservative, resulting in a higher rate of completely empty predictions for complex articles.

Based on these findings, we conclude that all four models trained on SYNTH-ITA fail on real DICE data due to identical **core issues**: insufficient exposure to the complex syntactic structures and narrative flow of real news articles, lack of training on nuanced role disambiguation in multi-participant narratives with complex coreference chains, inability to distinguish central crime elements (e.g., the stolen object) from incidental contextual mentions.

The failures are not primarily due to unseen vocabulary but rather to **unseen syntactic and discourse patterns** that require deeper understanding of narrative structure, coreference resolution, and pragmatic role assignment. While DS and Q14B are slightly more robust (exhibiting fewer complete failures), all models share the same fundamental limitations. The consistent failure patterns across four different underlying LLMs suggest that the issue lies in the **training data distribution** (i.e., the synthetic generation process) rather than in the specific model architecture.

5.2.4. Comparison with multilingual models

We conducted additional experiments with two widely-used multilingual encoders: **Multilingual BERT (mBERT)** (Devlin et al., 2019) and **XLM-RoBERTa (XLM-R)** (Conneau et al., 2020). Both models were fine-tuned for the QA-EE formulation using the same training procedure as our monolingual BERT and ELECTRA models in Section 5.1.

We decided to evaluate these models under two training regimes: (1) fine-tuning on the real $DICE_{200}$ dataset, and (2) fine-tuning on the large-scale synthetic $SYN2_{LL3\text{-}8B}$ dataset. Performance was measured on the $DICE_{406}$ test set, with results presented in Table 9.

Analysis of Results. The performance of multilingual models reveals important distinctions in their ability to leverage synthetic data compared to their monolingual counterparts from Table 7. When trained on the same $DICE_{200}$ real data, both mBERT and XLM-R achieve lower EM and PM F1 scores (approximately 40.2–40.3 and 48.9–49.9, respectively) than the monolingual Italian BERT (45.0 and 53.8) and ELECTRA (39.7 and 48.3). This aligns with expectations, as monolingual models typically outperform multilingual ones on language-specific tasks when sufficient in-language training data is available.

However, the results diverge significantly when the models are trained on synthetic data. **XLM-R** demonstrates a strong capacity to benefit from the SYN2 dataset, achieving an EM F1 of 44.1% and a PM F1 of 53.7%. These scores are competitive with the best monolingual models trained on synthetic data (e.g., BERT on $SYN2_{Q7B}$: EM F1 43.8, PM F1 54.6) and substantially outperform its

Table 9
Performance of Multilingual QA-EE Models on $DICE_{406}$ Test Set.

Model	Training data	EM (%)			PM (%)		
		P	R	F1	P	R	F1
mBERT	$DICE_{200}$	43.5	37.3	40.2	55.2	43.8	48.9
	$SYN2_{LL3\text{-}8B}$	3.3	49.5	6.2	15.4	57.5	24.3
XLM-R	$DICE_{200}$	42.0	38.6	40.3	54.5	46.0	49.9
	$SYN2_{LL3\text{-}8B}$	45.4	42.9	44.1	57.4	50.5	53.7

Table 10
TEST 2 - QA-EE and NER-EE performance on 80 manually annotated synthetic news from $SYN2_{LL3\text{-}8B}$ (i.e., $SYN2_{manual\text{-}80}$). The best scores for each model type are shown in bold, while the second-best scores are underlined.

Model	Training dataset	EM (%)			PM (%)		
		P	R	F1	P	R	F1
BERT	$MSQA_{ita}$	37.1	2.4	4.5	44.7	4.7	8.4
	$DICE_{10}$	23.7	1.3	2.5	23.7	1.3	2.5
	$DICE_{80}$	96.3	86.2	91.0	95.4	87.4	91.2
	$DICE_{200}$	96.2	93.2	94.7	96.5	94.3	95.4
	$SYN2_{LL3\text{-}8B}$	98.6	98.1	**98.4**	**98.0**	97.5	97.7
	$SYN2_{Q7B}$	98.6	**98.2**	**98.4**	**98.0**	**97.6**	**97.8**
	$SYN2_{Q14B}$	98.5	**98.2**	**98.4**	**98.0**	**97.6**	**97.8**
	$SYN2_{DS}$	**98.8**	98.0	97.7	**98.0**	97.4	97.7
ELECTRA	$MSQA_{ita}$	40.2	2.7	5.0	50.3	5.9	10.6
	$DICE_{10}$	23.7	1.3	2.5	23.7	1.3	2.5
	$DICE_{80}$	92.6	86.6	89.5	92.2	87.5	89.8
	$DICE_{200}$	96.4	91.1	93.7	96.6	91.9	94.2
	$SYN2_{LL3\text{-}8B}$	**98.7**	98.2	98.4	**98.0**	97.6	**97.8**
	$SYN2_{Q7B}$	98.5	98.3	**98.5**	**98.0**	97.6	**97.8**
	$SYN2_{Q14B}$	98.6	**98.4**	**98.5**	**98.0**	**97.7**	**97.8**
	$SYN2_{DS}$	**98.7**	98.3	**98.5**	**98.0**	97.6	**97.8**
SpacyNER	$DICE_{10}$	57.1	30.2	39.5	58.8	31.6	41.1
	$DICE_{80}$	59.5	41.0	48.5	64.9	48.7	55.6
	$DICE_{200}$	66.5	54.3	59.8	72.4	60.7	66.0
	$SYN2_{LL3\text{-}8B}$	**86.7**	76.8	81.5	**89.4**	81.4	**85.2**
	$SYN2_{Q7B}$	83.4	70.3	76.3	85.8	75.1	80.1
	$SYN2_{Q14B}$	82.6	**80.6**	**81.6**	85.6	**83.6**	84.6
	$SYN2_{DS}$	84.4	77.0	80.5	87.4	80.8	84.0

own performance when trained on only 200 real articles. This suggests that XLM-R's robust cross-lingual pretraining enables it to effectively generalize from the patterns present in the synthetic Italian data.

In contrast, **mBERT** performs very poorly when trained on $SYN2_{LL3\text{-}8B}$, with an EM F1 of only 6.2%. The extremely low precision (3.3%) indicates a severe issue with hallucination or incorrect span extraction, despite a relatively high recall (49.5%). This result highlights a critical weakness of mBERT in handling synthetically generated data for this specific task, possibly due to its smaller vocabulary and less effective cross-lingual representations compared to XLM-R.

Implications. This comparison underscores that the utility of synthetic data is not uniform across all model architectures. While XLM-R shows a promising ability to leverage large-scale synthetic data as effectively as monolingual models, mBERT fails to do so. This finding reinforces the importance of model selection when using synthetic data for training. Furthermore, it demonstrates that SYN2 can successfully augment training for certain state-of-the-art multilingual models, potentially enabling effective event extraction in Italian without requiring extensive monolingual pretraining.

5.3. Test 2 - models' evaluation on synthetic data

This test measures performance on a small portion of a synthetic dataset, but with reliable manual annotations. While in Test 1 we can see how models trained on the synthetic data perform on the real data, here, in Test 2, we see the performance of the models trained on the real data and evaluated on the synthetic data. In particular, we compare the models'performance when tested on 80 manually annotated news from $SYN2_{LL3\text{-}8B}$ (i.e., $SYN2_{manual\text{-}80}$). For models trained on $SYN2_{LL3\text{-}8B}$, Test 2 evaluates the internal consistency of synthetic data (validation of the dataset itself). Moreover, it is important to consider that similar news items may also be present in the other SYN2 variants, as they were generated from the same JSON sources as $SYN2_{LL3\text{-}8B}$.

The results of this test are shown in Table 10. Among the QA-EE models, both BERT and ELECTRA achieve their best performance when trained on synthetic datasets. These models consistently achieve EM F1 scores of 98.4% and PM F1 scores of 97.8%, with ELECTRA slightly edging out BERT in several cases. However, QA-EE models achieve excellent performance even when trained on a substantial number of DICE news articles. In particular, BERT trained on $DICE_{200}$ achieves 94.7% F1 score for EM and 95.4% for PM.

Table 11
Event argument extraction performance on real AC_{406} news: QA-EE vs NER-EE models trained on various datasets. Metrics include precision (P), recall (R) and F1 score. The best scores for each model type are shown in bold, while the second-best scores are underlined.

Model	Training data	EM (%)			PM (%)		
		P	R	F1	P	R	F1
ELECTRA	AC_{10}	–	–	–	–	–	–
	AC_{80}	<u>39.7</u>	<u>49.0</u>	<u>43.9</u>	<u>62.2</u>	<u>57.4</u>	<u>59.7</u>
	AC_{200}	**46.9**	**56.4**	**51.2**	**67.9**	**66.6**	**67.3**
	SYNAC	33.9	37.5	35.6	52.4	45.3	48.6
ELECTRA	AC_{10}	–	–	–	–	–	–
	AC_{80}	<u>35.5</u>	<u>50.1</u>	<u>41.6</u>	<u>57.4</u>	<u>62.4</u>	<u>59.8</u>
	AC_{200}	**53.3**	**59.8**	**57.8**	**70.2**	**72.9**	**71.5**
	SYNAC	33.5	38.1	35.6	51.9	46.7	49.2
SpacyNER	AC_{10}	–	–	–	–	–	–
	AC_{80}	<u>44.9</u>	<u>48.5</u>	<u>46.7</u>	<u>54.5</u>	<u>51.4</u>	<u>52.9</u>
	AC_{200}	**49.0**	**53.5**	**51.1**	**60.0**	**57.5**	**58.7**
	SYNAC	33.6	41.1	37.0	48.6	48.1	48.3

SpacyNER achieves a relatively high score when trained on synthetic datasets, reaching a PM F1 score of 85.2% when trained on $SYN2_{LL3\text{-}8B}$. However, unlike QA-EE models, the score remains more limited on datasets composed of DICE news articles, including $DICE_{200}$.

As can be seen, models trained on synthetic data in Test 2 have very high performances, and this might highlight the risk of overfitting. On the other hand, we know that there might be systematic bias in the synthetic dataset (e.g., repetitive patterns, lack of linguistic variance). Therefore, the high F1 scores would not reflect practical usefulness, but only the internal consistency of the synthetic datasets.

On the other hand, we can observe that the stability of the models trained on manually annotated real data (SpacyNER model trained on $DICE_{200}$ still reach around 60% as in Test 1) confirms that quality of training data is, in some cases, better than a huge amount of dataset to train (even a small but well-annotated dataset is preferable to a large synthetic one).

6. Cross-domain generalization: A preliminary case study on air crash events

To assess the generalizability of our proposed methodology beyond the theft-related crime domain, we conducted an additional experiment on a new domain: English news articles about air crash events. This domain differs substantially in topic, language, event structure, and linguistic characteristics, providing a robust test for our methodology. For this purpose, we used the DocEE dataset (Tong et al., 2022), a large-scale benchmark for document-level event extraction in English that includes a wide range of event types (Armed Conflict, Air Crash, Earthquakes, SportsCompetition, etc.). We focused on the "Air Crash" event type within DocEE, which comprises 1785 documents annotated with the following labels: *Date*, *Aircraft Agency*, *Scheduled Landing Place*, *Location*, *Flight Number*, and *Crew*.

Unlike the application on SYNTH-ITA, we did not repeat Phase 1 (LLM validation and selection). Instead, we directly used LL3-8B as the content generator, leveraging the fact that it was already evaluated in the original domain and exhibited good performance in terms of structural alignment and textual quality. Our efforts focused entirely on adapting Phase 2 to the new domain.

Prerequisites: Event Schema and Real Annotated Data: We split the annotated data into five disjoint subsets: 858 articles used for JSON generation (AC_{858}), 406 for testing (AC_{406}), and 10, 80, and 200 articles to define respectively AC_{10}, AC_{80}, and AC_{200}. The split was performed so that each subset contains at least one occurrence of every possible label. Unlike DICE, DocEE was annotated via crowdsourcing without clear annotation guidelines, leading to lower precision than expert-labeled data.

Step 2.1 - Entity Description Generation: Since DocEE did not define a annotation schema, we extrapolate label definitions from the annotated news.

Step 2.2 - Randomized JSON Generation: We extracted all annotated spans for each label from AC_{858}, defining a list for each label. These lists were expanded using LLMs and manually validated to create rich, realistic entity lists for JSON generation. Using our randomized JSON generation process with the curated entity lists, we created 38,000 JSON fictitious annotations.

Step 2.3 - Iterative Generation-Evaluation Loop of Synthetic News: Building on previously defined label specifications, we designed a targeted generation prompt that takes a JSON input and produces a synthetic air crash news report. This prompt was constructed using four real examples from AC_{858}, selected to ensure coverage of all six labels. The prompt was initially tested with LL3-8B to create a set of news articles, which were manually examined to evaluate quality and adherence to the annotations. Several issues were identified, leading to further refinement of the prompt. The revised version was then applied with LL3-8B to the 38,000 JSON inputs, generating the synthetic air crash dataset.

Step 2.4 - Post-Processing, Evaluation and Released: Unlike the post-processing pipeline used for SYNTH-ITA, we did not implement a rule-based algorithm that searches for alternative formulations or updates JSON spans to match the text. Instead, we applied a strict validation step that checks for the presence of each JSON span in the generated text. If any span was missing, the news was discarded. This conservative filtering ensures that all remaining news are perfectly aligned with their source JSON, but

Table 12
Label-wise F1 scores on PM for QA-EE and NER-EE formulations on AC_{406} (numbers inside brackets exclude news with empty annotations for that label from the ground-truth). The best results per label-model pair are shown in bold, while the second-best scores are underlined.

		Date	Aircraft agency	Scheduled landing place	Location	Flight number	Crew
				Label statistics			
	#annotated spans	396	343	306	272	351	214
	% news with annotations	96	83	73	64	70	33
	Training dataset			**F1-score (%)**			
BERT	AC_{10}	–	–	–	–	–	–
	AC_{80}	**90.4** (**91.6**)	67.7 (69.2)	30.2 (18.6)	47.2 (43.9)	65.3 (**72.4**)	–
	AC_{200}	89.0 (90.5)	**76.6** (**81.7**)	**63.9** (**66.0**)	**59.5** (**57.6**)	**67.8** (71.8)	**51.4** (**36.9**)
	SYNAC	87.3 (89.0)	47.6 (40.2)	30.7 (32.5)	46.9 (34.0)	29.2 (13.9)	48.9 (24.8)
ELECTRA	AC_{10}	–	–	–	–	–	–
	AC_{80}	89.6 (91.5)	70.1 (69.9)	48.7 (56.5)	44.7 (62.8)	57.3 (69.9)	–
	AC_{200}	**90.6** (**91.9**)	**81.6** (**88.3**)	**74.9** (**74.8**)	**67.8** (**67.0**)	**65.8** (**70.5**)	51.3 (**49.1**)
	SYNAC	86.0 (87.1)	55.9 (49.4)	30.1 (34.6)	48.9 (39.9)	31.3 (18.9)	42.2 (33.7)
SpacyNER	AC_{10}	–	–	–	–	–	–
	AC_{80}	68.7 (42.2)	43.7 (32.4)	–	–	73.9 (66.7)	–
	AC_{200}	68.0 (68.3)	**65.4** (59.9)	**41.2** (23.1)	–	**76.4** (**71.5**)	–
	SYNAC	**74.6** (**76.4**)	50.0 (43.1)	36.0 (**32.1**)	**40.1** (**12.5**)	30.3 (8.0)	**50.6** (**40.1**)

it also reduces the acceptance rate. After this filtering, we obtained the final synthetic dataset, SynAC, comprising 3182 articles. We evaluate SYNAC by comparing it with AC_{406} across textual quality and diversity dimensions. First, AC_{406} contains substantially longer texts (average length 3948 tokens) compared to SYNAC (1298 tokens), along with a much larger vocabulary (322 vs. 122 unique tokens). This difference is also reflected in lexical diversity, where AC_{406} consistently achieves higher values across all measures, MTLD (73.2 vs. 59.8), HD-D (0.81 vs. 0.77), and MATTR (0.78 vs. 0.75). Self-BLEU increases dramatically from 0.01 in AC_{406} to 0.54 in SYNAC, indicating a substantial rise in redundancy in synthetic data. These results show that, although SYNAC approximates some statistical properties of the original corpus, it exhibits clear reductions in linguistic richness and variability.

Downstream evaluation: We evaluated the effectiveness of SYNAC by training QA-EE (based on BERT[12] and ELECTRA[13]) and NER-EE (based on English SpacyNER) models on the SYNAC dataset and comparing their performance against the same models trained on few-shot real news datasets (AC_{10}, AC_{80}, and AC_{200}). Table 11 presents the overall extraction performance on AC_{406} while Table 12 reports label-specific F1 scores for PM. All models trained on the extremely small dataset AC_{10} fail to generalize and produce only empty outputs, highlighting the limitations of using very limited real annotations. Training the model on SYNAC does not reach the performance achieved by models trained on AC_{80}, but the results are relatively close. In contrast, the gap with models trained on AC_{200} remains substantial.

From Table 12, models AC_{80} achieves relatively strong performance in the QA-EE setting, where it generally outperforms SYNAC. However, this advantage does not transfer to the NER-EE models where AC_{80} exhibits unstable behavior, with less frequent labels receiving empty predictions. A similar pattern is observed for AC_{200}, which still yields empty predictions for certain labels under NER-EE. On the other hand, SYNAC shows a more robust and consistent behavior in the NER formulation. Although its overall performance may be lower in some cases compared to QA-based approaches, it does not suffer from the "all-empty" prediction issue observed in AC-based configurations. This indicates that SYNAC ensures more reliable coverage across labels, particularly in the NER-EE formulation. For the *Crew* label, which appears in only about 33% of AC_{406}, the models trained on AC_{80} are unable to correctly predict any instances. This behavior is likely due to the limited exposure to this label during training, as it occurs in only a small number of the annotated articles, preventing the models from learning patterns that generalize to the remaining documents. In this scenario, having access to a larger manually annotated dataset, such as AC_{200}, becomes necessary. Alternatively, the support provided by models trained on synthetic datasets can help mitigate this limitation.

These findings suggest that a synthetic dataset such as SYNAC can be a useful alternative when the goal is to approximate the performance obtainable with a moderately sized manually annotated dataset without the cost of annotating a large number of news articles. Moreover, synthetic data can provide better coverage of underrepresented labels, mitigating the effects of label sparsity in small manually annotated datasets and enabling models to learn patterns associated with infrequent arguments.

7. Discussion

The SYNTH-ITA corpus successfully addresses a critical resource gap for Italian NLP, providing the first medium-scale dataset with fine-grained event annotations for theft events.

[12] https://huggingface.co/google-bert/bert-base-uncased
[13] https://huggingface.co/google/electra-base-discriminator

7.1. Theoretical and practical implications

Our downstream evaluation (Section 5) reveals a more nuanced picture than simple resource augmentation. The results demonstrate a clear *task- and entity-dependent performance gap*, highlighting that the value and limitations of synthetic data are contingent on specific application scenarios. While carefully designed synthetic datasets like SYNTH-ITA show significant potential for building accurate models in target domains, their effectiveness is not uniform and requires critical examination of underlying trade-offs.

A deeper analysis of the performance divergence between QA-EE and NER-EE models reveals fundamental differences in how these paradigms represent task information. Traditional NER models based on token-level sequence labeling represent roles as discrete labels assigned to tokens, lacking explicit semantic grounding and encouraging the model to learn statistical token-label associations. Empirical studies have shown that NER systems rely heavily on surface forms seen during training and generalize poorly to unseen entities or distribution shifts (Augenstein et al., 2017; Ma et al., 2023). Consequently, when trained on synthetic data that does not perfectly mirror the complexity of real contexts, NER models suffer performance degradation (Kamath & Vajjala, 2025). In contrast, QA-based EE formulations express argument roles as natural language queries that explicitly encode the semantic definition of each role. This query conditioning injects label semantics directly into the input representation, enabling the model to align contextual evidence with a role description rather than relying solely on surface token-label correlations. Prior work demonstrates that span extraction problems such as NER and EE benefit from this additional task signal when cast into QA-style tasks (Arora & Park, 2023; Du & Cardie, 2020; Yang et al., 2021).

- **Task-Dependent Effectiveness** Our findings demonstrate that the utility of synthetic data is not uniform but highly dependent on task formulation. QA-EE models, which frame extraction as information retrieval from context, benefit significantly from SYNTH-ITA's structural consistency and scale, achieving performance comparable to models trained on 200 manually annotated articles. In contrast, NER-EE models, which require precise boundary recognition, show an 18% performance drop. The cross-domain experiment on English air crash events further reinforces this theoretical picture. Even in a new domain, QA-EE models trained on synthetic data (SYNAC) outperformed models trained on very small real datasets, whereas NER-EE models struggled unless provided with a substantial number of real examples. The robustness of QA-EE to domain shifts appears to stem from the same semantic grounding mechanism, suggesting that the advantage is not limited to the original crime domain.
- **Entity-Specific Limitations Reveal Generation Gaps** The uneven performance across event roles further delimits SYNTH-ITA's applicability. While the pipeline reliably generates coherent content for high-frequency, concrete roles (LOC, OBJ, PAR), it falls short on person-related entities (AUT/AUTG, VIC/VICG), which require more nuanced contextual expression. This indicates that the current methodology, while controlled, cannot fully replicate the linguistic diversity and pragmatic knowledge associated with human participants in real crime narratives.
- **Synthetic Data As Complementary Resources** Results collectively suggest that SYNTH-ITA should not be viewed as a complete substitute for real annotated data, but rather as a complementary resource with specific strengths. Its primary value lies in: (1) *pre-training* or *large-scale augmentation* to build robust base models, (2) providing *schema-aligned examples* for tasks that prioritize information retrieval over precise boundary detection, and (3) *enabling research* in low-resource scenarios where any annotated data is scarce. However, for production systems requiring high precision on complex entities, a hybrid approach combining synthetic scale with targeted real annotations remains preferable.

Preliminary experiments on a different domain (air crash events) further support this view, showing that synthetic data's utility remains task- and domain-dependent.

7.2. Limitations and boundary conditions

The following limitations, some inherent to synthetic data generation and others specific to our implementation, define the boundary conditions within which SYNTH-ITA and our methodology are effective. These should inform potential users' decisions about when and how to employ such synthetic resources.

- **The Scalability-Authenticity Trade-off** As noted, the synthetic datasets achieve high structural alignment at the cost of reduced linguistic authenticity. The lower MTLD values and the higher cross-dataset divergence (JSD) indicate that the generated texts are less lexically diverse than real news. Future work could explore more sophisticated prompting strategies (e.g., incorporating discourse-level constraints) or employ retrieval-augmented generation to ground the text in authentic exemplars.
- **Annotation Consistency** Although post-processing steps significantly improved alignment, some inconsistencies may persist in the structured annotations, particularly when dealing with multi-word expressions, coreference resolution, and label assignment. Manual review and further refinement may be necessary to enhance annotation accuracy.
- **Bias Assessment** A limitation of this study lies in the analysis of potential biases in the generated news articles. While we implemented fairness safeguards at the input level by randomizing gender and nationality attributes in the synthetic JSONs and ensuring balanced representation in the prompts, a significant limitation is that we did not systematically evaluate whether the language model preserved this balance in the generated text. The synthetic dataset's demographic fairness was verified only at the input level, leaving open the possibility that the model introduced or amplified biases in the generated text. For instance,

the LLM might disproportionately associate certain nationalities with criminality or reinforce gender stereotypes despite neutral input conditions. Future work should incorporate post-generation bias detection to quantify deviations between input prompts and generated text.

- **Single-event assumption** Our methodology assumes that each news article contains exactly one theft event. This simplifies the event extraction task but does not reflect real-world scenarios where texts may contain zero or multiple events. This design choice is intentional and reflects a modular pipeline: first, documents that describe no events are discarded through a classification step; for documents containing multiple events, within-document event coreference resolution is applied to isolate text segments corresponding to individual events. After this preprocessing step, each document consists of text describing a single event, making it suitable for our methodology. As such, SYNTH-ITA focuses on isolating and evaluating fine-grained argument extraction under controlled conditions. Nevertheless, this assumption limits the dataset's direct applicability to end-to-end event extraction settings where such preprocessing is not performed.
- **Cross-Domain Challenges** The adaptation of our methodology on a different domain (air crash events in English) showed that synthetic data can improve over very small real datasets, but it did not reach the performance of larger real corpora. More importantly, it highlighted that successful application relies on:
  - A well-defined annotation schema with clear labeling guidelines;
  - Deep familiarity with the target domain to enable effective prompt refinement;
  - Domain-specific post-processing rules, especially for long spans and multi-word expressions.

  Without these, the methodology may produce lower-quality outputs or require extensive manual filtering.
- **Broader Applications** Beyond event extraction, synthetic datasets generated with this pipeline could be used for other tasks such as summarization, question answering. Investigating such applications would broaden the impact of the resource.

7.3. Methodological framework adaptation

Our two-phase, document-level generation pipeline offers a reusable framework for creating synthetic datasets from structured annotations in other domains. However, successful adaptation depends on several critical conditions:

- **Prerequisite: Well-Defined Schema** The approach requires a clear, consistent annotation schema. with ambiguous or highly variable event structures may not benefit equally from this methodology.
- **Need for Real-World Anchoring** A small set of high-quality real documents (100–200) is essential for prompt refinement and validation. Without this anchor, the generated text may diverge significantly from domain-appropriate language.This small but representative sample serves both as a reference for linguistic fidelity and as a basis for evaluating structural alignment.
- **Domain Suitability** The methodology works best for domains with relatively predictable narrative structures (e.g., news reports, clinical notes). It may be less effective for highly creative or conversational domains where narrative coherence is more complex. Moreover, our current framework assumes each document describes a single event; for domains where multiple events co-occur, additional preprocessing such as event segmentation is required.
- **Risk Awareness** Adapting the framework requires careful attention to: (1) *Bias amplification* - the LLM may introduce biases not present in the structured input; (2) *Overfitting to synthetic patterns* - models may learn artifacts of the generation process rather than generalizable features; (3) *Scalability-authenticity trade-off* - larger generated datasets may exhibit reduced lexical diversity and shorter texts, as observed in SYNTH-ITA. For applications where stylistic authenticity is critical, synthetic data may need to be complemented with techniques such as style transfer or human-in-the-loop refinement.

These conditions define the *practical applicability* of our methodology: it is most valuable for creating training data in structured, predictable domains where real annotations are scarce, but requires careful validation and likely hybrid approaches for deployment in production systems.

8. Conclusion

This paper introduced **SYNTH-ITA**, the first collection of medium-scale synthetic datasets for fine-grained theft event extraction in Italian. Our work addresses the critical resource gap in Italian NLP by proposing a novel two-phase, document-level synthetic data generation pipeline that leverages Large Language Models (LLMs) under a controlled, schema-driven framework.

The proposed methodology ensures structural fidelity, narrative coherence, and bias mitigation through explicit alignment mechanisms, rule-based post-processing, and iterative expert validation. The resulting datasets, each containing thousands of high-fidelity Italian theft news articles with fine-grained event annotations, demonstrate high textual quality, lexical diversity, and strong alignment with real crime reports. Downstream evaluations reveal that models trained on SYNTH-ITA, especially those using a Question Answering formulation for event extraction, achieve performance comparable to or better than models trained on 200 manually annotated articles, with improvements of up to 14% in F1 score for ELECTRA-based models. However, NER-based models show an 18% performance drop, highlighting the task-dependent utility of synthetic data.

SYNTH-ITA is released as an open resource, providing a scalable and complementary alternative to manual annotation for low-resource NLP scenarios, while also serving as a benchmark for document-level event extraction in Italian.

The proposed methodology and SYNTH-ITA dataset open several promising research directions. Future work should focus on: (1) expanding to new domains and languages; (2) enhancing textual quality through advanced prompting and discourse modeling; (3)

implementing systematic post-generation bias analysis; (4) relaxing the single-event assumption to handle complex event structures; (5) exploring hybrid strategies that combine synthetic and real data with human feedback; and (6) evaluating applicability to broader NLP tasks such as summarization, translation, and multi-modal reasoning. In parallel, the development of semi-automated tools to assist in schema extraction and prompt design could reduce the manual effort required for adapting the pipeline to new domains.

CRediT authorship contribution statement

Giovanni Bonisoli: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Data curation, Conceptualization. **Federica Rollo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Laura Po:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used chatGPT and DeepSeek to improve the readability and language of the manuscript. After using these tools, the authors reviewed and edited the content as needed and assume full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the newspaper “Gazzetta di Modena” that allows the collection and distribution of the corpus under a CC BY-NC-SA 4.0 license.

Data availability

The datasets and the source code of our methodology are available at https://github.com/federicarollo/SYNTH-ITA.

References

Arora, J., & Park, Y. (2023). Split-NER: Named entity recognition via two question-answering-based classifications. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 416–426). Toronto, Canada: Association for Computational Linguistics.

Association, A. P., General, P. M., & Applied (1944). Studies in language behavior: I. A program of research / by Wendell Johnson. In *Psychological monographs*: *vol. 56*, (2), (pp. 1–15). District of Columbia: American Psychological Association, 1944.

Augenstein, I., Derczynski, L., & Bontcheva, K. (2017). Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, *44*, 61–83.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., & Mazzoleni, M. (2004). Introducing the la repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In *Proceedings of the fourth international conference on language resources and evaluation* (pp. 1771–1774). Lisbon, Portugal: European Language Resources Association (ELRA).

Bonifacio, L., Abonizio, H., Fadaee, M., & Nogueira, R. (2022). Inpars: Data augmentation for information retrieval using large language models.

Bonisoli, G., Di Buono, M. P., Po, L., & Rollo, F. (2023). DICE: a dataset of Italian crime event news. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval* (pp. 2985–2995). New York, NY, USA: Association for Computing Machinery.

Bonisoli, G., Vilares, D., Rollo, F., & Po, L. (2025). Document-level event extraction from Italian crime news using minimal data. *Knowledge-Based Systems*, *317*, Article 113386.

Chen, S. F., & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *10*(4), 359–393.

Chen, J., Tam, D., Raffel, C., Bansal, M., & Yang, D. (2023). An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the Association for Computational Linguistics*, *11*, 191–211.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., ... Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*, 70:1–70:53.

Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th international conference on learning representations* (pp. 1–18). OpenReview.net.

Cloud, A., & Team, Q. (2024). Qwen2.5 technical report.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics.

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100.

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Zeng, F., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., & Li, X. (2025). AugGPT: Leveraging ChatGPT for text data augmentation. *IEEE Transactions on Big Data*, *11*(3), 907–918.

de Rosa, G. H., & Papa, J. P. (2021). A survey on text generation using generative adversarial networks. *Pattern Recognition*, *119*, Article 108098.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., & Joty, S. (2024). Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: ACL 2024* (pp. 1679–1705). Bangkok, Thailand: Association for Computational Linguistics.

Du, X., & Cardie, C. (2020). Event extraction by answering (almost) natural questions. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 671–683). Online: Association for Computational Linguistics.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the association for computational linguistics* (pp. 968–988). Online: Association for Computational Linguistics.

Feng, Y., Li, L., Qin, X., & Zhang, B. (2025). Improving event representation learning via generating and utilizing synthetic data. *Information Processing & Management*, *62*(4), Article 104083.

Gao, J., Yu, C., Wang, W., Zhao, H., & Xu, R. (2022). Mask-then-fill: A flexible and effective data augmentation framework for event extraction. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the association for computational linguistics* (pp. 4537–4544). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Gatto, J., Sharif, O., Seegmiller, P., & Preum, S. M. (2025). Document-level event-argument data augmentation for challenging role types. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 25109–25131). Vienna, Austria: Association for Computational Linguistics.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in neural information processing systems*: *vol. 27*, (pp. 2672–2680). Curran Associates, Inc..

Grattafiori, A., et al. (2024). The llama 3 herd of models.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems*: *vol. 33*, (pp. 6840–6851). Curran Associates, Inc..

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B. CoRR, abs/2310.06825.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M., Stock, P., Subramanian, S., Yang, S., ... Sayed, W. E. (2024). Mixtral of experts. CoRR, abs/2401.04088.

Jin, X., & Ji, H. (2024). Schema-based data augmentation for event extraction. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation* (pp. 14382–14392). Torino, Italia: ELRA and ICCL.

Josifoski, M., Sakota, M., Peyrard, M., & West, R. (2023). Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 1555–1574). Singapore: Association for Computational Linguistics.

Kamath, G., & Vajjala, S. (2025). Does synthetic data help named entity recognition for low-resource languages? In K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, & D. P. Singh (Eds.), *Proceedings of the 14th international joint conference on natural language processing and the 4th conference of the Asia-Pacific chapter of the association for computational linguistics* (pp. 159–167). Mumbai, India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Kingma, D. P., & Welling, M. (2022). Auto-encoding variational Bayes.

Krippendorff, K. (2006). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, *30*(3), 411–433.

Li, Y., Ding, K., Wang, J., & Lee, K. (2024). Empowering large language models for textual data augmentation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: ACL 2024* (pp. 12734–12751). Bangkok, Thailand: Association for Computational Linguistics.

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 110–119). San Diego, California: Association for Computational Linguistics.

Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *AI Open*, *3*, 71–90.

Li, H., Tomko, M., Vasardani, M., & Baldwin, T. (2022). MultiSpanQA: A dataset for multi-span question answering. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1250–1260). Seattle, United States: Association for Computational Linguistics.

Li, Y., Zhang, Y., Du, Z., & Guo, Z. (2025). Large language model data augmentation for text-pair classification tasks. In *Proceedings of the 2024 13th international conference on computing and pattern recognition* (pp. 427–433). New York, NY, USA: Association for Computing Machinery.

Liu, J., Chen, Y., & Xu, J. (2021). Machine reading comprehension as data augmentation: A case study on implicit event argument extraction. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 2716–2725). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Liu, J., Chen, Y., & Xu, J. (2022). Mrcaug: Data augmentation via machine reading comprehension for document-level event argument extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 3160–3172.

Liu, H., Teng, Z., Cui, L., Zhang, C., Zhou, Q., & Zhang, Y. (2023). LogiCoT: Logical chain-of-thought instruction-tuning.

Lu, J., Henchion, M., & Mac Namee, B. (2020). Diverging divergences: Examining variants of jensen Shannon divergence for corpus comparison tasks. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 6740–6744). Marseille, France: European Language Resources Association.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A., & Pirrelli, V. (2014). The PAISÀ corpus of Italian web texts. In *Proceedings of the 9th web as corpus workshop (wac-9)* (pp. 36–43). Gothenburg, Sweden: Association for Computational Linguistics.

Ma, R., Wang, X., Zhou, X., Zhang, Q., & Huang, X. (2023). Towards building more robust NER datasets: An empirical study on NER dataset bias from a dataset difficulty view. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 4616–4630). Singapore: Association for Computational Linguistics.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392.

Minard, A.-L., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., & van Son, C. (2016). MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the tenth international conference on language resources and evaluation* (pp. 4417–4422). Portorož, Slovenia: European Language Resources Association (ELRA).

Nadăş, M., Dioşan, L., & Tomescu, A. (2025). Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, *13*, 134615–134633.

Piedboeuf, F., & Langlais, P. (2022). Effective data augmentation for sentence classification using one VAE per class. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th international conference on computational linguistics* (pp. 3454–3464). Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Polignano, M., Basile, P., & Semeraro, G. (2024). Advanced natural-based interaction for the italian language: LLaMAntino-3-ANITA.

Puduppully, R., & Lapata, M. (2021). Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, *9*, 510–527.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics.

Rovera, M. (2024). EventNet-ITA: Italian frame parsing for events. In Y. Bizzoni, S. Degaetano-Ortlieb, A. Kazantseva, & S. Szpakowicz (Eds.), *Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature* (pp. 77–90). St. Julians, Malta: Association for Computational Linguistics.

Salkar, N., Trikalinos, T., Wallace, B., & Nenkova, A. (2022). Self-repetition in abstractive neural summarizers. In Y. He, H. Ji, S. Li, Y. Liu, & C.-H. Chang (Eds.), *Proceedings of the 2nd conference of the Asia-Pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing (volume 2: short papers)* (pp. 341–350). Online only: Association for Computational Linguistics.

Team, D.-A., Liu, A., Feng, B., et al. (2024). DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model.

Tian, X., Guo, Y., Ge, B., Yuan, X., Zhang, H., Yang, Y., Ke, W., & Li, G. (2024). Agent-DA: Enhancing low-resource event extraction with collaborative multi-agent data augmentation. *Knowledge-Based Systems*, *305*, Article 112625.

Tonelli, S., Sprugnoli, R., & Moretti, G. (2019). Prendo la parola in questo consesso mondiale: A multi-genre 20th century corpus in the political domain. In *CEUR Workshop Proceedings*: *vol. 2481*.

Tong, M., Xu, B., Wang, S., Han, M., Cao, Y., Zhu, J., Chen, S., Hou, L., & Li, J. (2022). DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 3970–3982). Seattle, United States: Association for Computational Linguistics.

Wang, S., & Huang, L. (2024). Targeted augmentation for low-resource event extraction. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the association for computational linguistics* (pp. 4414–4428). Mexico City, Mexico: Association for Computational Linguistics.

Wang, B., Huang, H., Wei, X., Shi, G., Liu, X., Feng, C., Zhou, T., Wang, S., & Yin, D. (2023). Boosting event extraction with denoised structure-to-text augmentation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023* (pp. 11267–11281). Toronto, Canada: Association for Computational Linguistics.

Yang, P., Cong, X., Sun, Z., & Liu, X. (2021). Enhanced language representation with label knowledge for span extraction. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 4623–4635). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Ye, J., Li, C., Kong, L., & Yu, T. (2023). Generating data for symbolic language with large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 8418–8443). Singapore: Association for Computational Linguistics.

Ye, J., Xu, N., Wang, Y., Zhou, J., Zhang, Q., Gui, T., & Huang, X. (2024). LLM-DA: Data augmentation via large language models for few-shot named entity recognition.

Ye, G., Zhao, H., Zhang, Z., & Jiang, Z. (2025). UniDE: A multi-level and low-resource framework for automatic dialogue evaluation via LLM-based data augmentation and multitask learning. *Information Processing & Management*, *62*(3), Article 104035.

Yi, Q., Chen, X., Zhang, C., Zhou, Z., Zhu, L., & Kong, X. (2024). Diffusion models in text generation: a survey. *PeerJ Computer Science*, *10*, Article e1905.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *8th international conference on learning representations* (pp. 1–43). OpenReview.net.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., & Yu, Y. (2018). Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 1097–1100). New York, NY, USA: Association for Computing Machinery.