QA-Noun: Representing Nominal Semantics via Natural Language Question-Answer Pairs

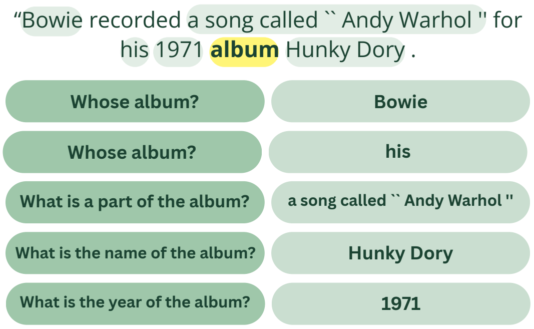
Maria Tseytlin[1,2] **Paul Roit**[2] **Omri Abend**[1] **Ido Dagan**[2] **Ayal Klein**[3]
[1]Hebrew University of Jerusalem [2]Bar-Ilan University
[3]Ariel University
maria.tseytlin@mail.huji.ac.il

Abstract

Decomposing sentences into fine-grained meaning units is increasingly used to model semantic alignment. While QA-based semantic approaches have shown effectiveness for representing predicate-argument relations, they have so far left noun-centered semantics largely unaddressed. We introduce QA-Noun, a QA-based framework for capturing noun-centered semantic relations. QA-Noun defines nine question templates that cover both explicit syntactical and implicit contextual roles for nouns, producing interpretable QA pairs that complement verbal QA-SRL. We release detailed guidelines, a dataset of over 2,000 annotated noun mentions, and a trained model integrated with QA-SRL to yield a unified decomposition of sentence meaning into individual, highly fine-grained, facts. Evaluation shows that QA-Noun achieves near-complete coverage of AMR's noun arguments while surfacing additional contextually implied relations, and that combining QA-Noun with QA-SRL yields over 130% higher granularity than recent fact-based decomposition methods such as FactScore and DecompScore. QA-Noun thus complements the broader QA-based semantic framework, forming a comprehensive and scalable approach to fine-grained semantic decomposition for cross-text alignment.

1 Introduction

Semantic representations of text, which decompose sentence meaning into smaller information units, are increasingly recognized as essential for modeling fine-grained alignment between texts. Such decomposition based alignments are beneficial for assessing the faithfulness of generated texts against a source of truth, highlighting which parts were preserved, omitted or hallucinated (Fan et al., 2023; Qiu et al., 2024; Chen et al., 2023), and they can also be used to guide the generative process with better content selection (Narayan et al., 2023; Zhang et al., 2025).



"Bowie recorded a song called `` Andy Warhol '' for his 1971 **album** Hunky Dory .

Whose album?	Bowie
Whose album?	his
What is a part of the album?	a song called `` Andy Warhol ''
What is the name of the album?	Hunky Dory
What is the year of the album?	1971

Figure 1: Example QA-Noun annotations for a single target noun (**highlighted**), explicating each "atomic" fact involving the noun as an individual QA pair.

To meet these needs, recent work has explored decomposing text into discrete meaning units, either in interrogative form (Deutsch et al., 2021; Durmus et al., 2020; Honovich et al., 2021) or as declarative "atomic facts" (Min et al., 2023; Wanner et al., 2024). These approaches, while effective for downstream evaluation, lack an underlying representation framework for systematically covering all fine-grained information units. To overcome these issues, some recent work has revisited traditional NLP semantic formalisms, such as Semantic Role Labeling (SRL) and Abstract Meaning Representation, which rely on formal semantic schemata to model meaning via labeled predicate–argument relations (Qiu et al., 2024; Fan et al., 2023).

An alternative semantic representation paradigm, based on natural-language question–answer (QA) pairs, was pioneered by **QA-SRL** (He et al., 2015). Instead of relying on formal role inventories, QA-SRL expresses predicate–argument relations through simple, intuitive questions (e.g., "Who did X?") answered with spans from the sentence. This format is easily interpretable for lay annotators, supports efficient crowdsourcing, scales across domains and languages, and aligns naturally with language models — making it well-suited for both

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, pages 2727–2741
December 20-24, 2025 ©2025 Association for Computational Linguistics

human and automated model-driven semantic annotation. Recently, this method was expanded to target deverbal nominalizations (Klein et al., 2020), discourse relations (Pyatkin et al., 2020) and adjectives (Pesahov et al., 2023). However, nominal relations expressed between a noun and its arguments have largely been overlooked.

This work introduces **QA-Noun**, a QA-based representation for nominal semantics that is both layman-attainable and semantically comprehensive — capturing not only explicit grammatical roles, but also implicit relations inferred from context. QA-Noun targets nine core semantic dimensions relevant to nouns, each instantiated through a corresponding question template (See Figure 1 for an illustrative example). This design enables interpretable, highly fine-grained decomposition into minimal facts involving the noun — where each QA pair corresponds to a single semantic relation. Notably, decomposition granularity is crucial for accurately modeling semantic alignment across texts since misalignments can occur for any minimal individual fact. For example, in Figure 1, each QA pair represents an atomic fact whose faithfulness to a source must be assessed independently.

We make the following contributions: (1) we extend the QA-based semantic paradigm to cover noun semantics, complementing verbal QA-SRL to enable an exhaustive, fine-grained decomposition of sentence meaning; (2) we design a novel annotation framework for nouns, capturing nine core semantic dimensions through interpretable question templates (§3); (3) we release a high-quality dataset of noun-centered QAs (§4) and assess its consistency and coverage (§5); (4) we develop and evaluate QA-Noun models and integrate them with QA-SRL into a unified decomposition tool (§6); and (5) we show that combining QA-Noun with QA-SRL yields a 130%–150% gain in semantic granularity over prior fact-based decomposition methods (§7).

2 Background

2.1 Semantic Representations of Nouns

Different semantic formalisms such as PropBank (Palmer et al., 2005) and NomBank (Meyers et al., 2004) have sought to capture some of the information that a text conveys within a structured scheme. NomBank annotates nominal predicates in the Penn Treebank with their arguments, assigning each a semantic role from a fixed inventory. For example, in the phrase *"higher **rate** of improvement"*, NomBank labels *improvement* as the `theme`, and *higher* as the `value` of the predicate **rate**. We use NomBank as a key reference in the design of QA-Noun: its argument structures informed our initial question template set and served as a basis for assessing coverage during the design phase.

A later and widely adopted framework is **Abstract Meaning Representation** (AMR; Banarescu et al., 2013), which encodes sentence-level semantics as rooted, labeled graphs. AMR extends beyond predicate–argument structures to capture a broader range of semantic relations, including nominal ones. Its role inventory combines the PropBank roleset for eventive predicates with general relations such as `:poss`, `:part`, and `:consist-of`, enabling the representation of non-eventive and relational nouns. Because of its broad scope and community adoption, AMR is a central comparison point for QA-Noun. In Section 5.2, we show that our question templates capture essentially all noun arguments annotated in AMR, and often surface additional, contextually implied relations.

NomBank and AMR represent key prior efforts in modeling nominal semantics but rely on formal role inventories and expert annotation, which limits scalability. **NounAtlas** (Navigli et al., 2024) expands nominal SRL coverage through large-scale automatic projection and clustering of argument structures within the traditional SRL paradigm. **TNE** (Elazar et al., 2022) instead leverages natural-language prepositions as intuitive relation labels to annotate noun phrase relations at scale. Neither, however, is designed to provide a predicate–argument representation of nouns in context as QA-Noun does.

2.2 Semantic QA Approach

To avoid formal, hard-to-scale role inventories, QA-SRL (He et al., 2015) introduced a natural language–based representation in which arguments and their roles are expressed as question–answer pairs. In this formulation, the question encodes the semantic role, while the answers identify the corresponding arguments. This format does not depend on predefined semantic role lexicons, can be explained to annotators with minimal training due to its use of natural language, and captures valuable implicit arguments that may not be explicit in syntax (Roit et al., 2020).

Building on QA-SRL, the paradigm has gradually expanded into a broader **QA-based semantics**

(QASem) framework. This includes extensions to deverbal nominalizations (Klein et al., 2020), adjectives (Pesahov et al., 2023), and discourse relations (Pyatkin et al., 2020), moving toward a unified question–answer representation for predicate–argument structure.

Beyond annotation, QASem has proven effective as a semantic decomposition layer for downstream tasks. Recent work has leveraged QA-based predicate–argument units for fine-grained cross-text alignment and evaluation: QAAlign (Brook Weiss et al., 2021) aligns information across texts via QA pairs; Roit et al. (2024) leverage the induced QA-SRL grammar to detect arguments across sentences; QAPyramid (Zhang et al., 2025) uses QASRL/QANom units for pyramid-style content selection evaluation in summarization; and Cattan et al. (2024) apply QA-SRL and QANom to localize factual inconsistencies in attributable text generation. Together, these studies demonstrate that QA-based predicate–argument representations provide an intuitive and fine-grained decomposition of meaning that supports evaluating faithfulness, information selection, and source attribution.

In this work, we extend this QA-based paradigm to cover a wide range of noun-centered semantic relations, complementing QA-SRL's verbal focus and completing a major step toward comprehensive QA-based semantic decomposition.

2.3 Granular Semantic Decomposition

A growing body of work in text generation and evaluation has highlighted the benefits of decomposing sentences into smaller, interpretable meaning units to enable precise cross-text alignment. The **QG-QA** framework exemplifies this trend by representing sentences as sets of question–answer pairs, supporting fine-grained evaluation of semantic overlap and faithfulness in summarization and factuality tasks (Eyal et al., 2019; Gavenavicius, 2020; Deutsch et al., 2021; Honovich et al., 2021; Durmus et al., 2020).

In a similar spirit, recent systems such as **FactScore** (Min et al., 2023), (Zhu et al., 2024) and **DecompScore** (Wanner et al., 2024) decompose sentences into sets of "atomic" natural-language facts for factuality and content selection evaluation. These approaches demonstrate that finer-grained decompositions yield more accurate and interpretable measures of semantic consistency. However, they typically rely on prompting large language models without a defined schema, which makes the granularity and coverage of the resulting "atomic facts" difficult to control in a principled way.

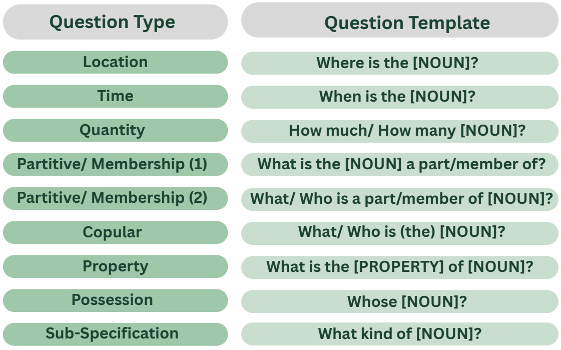
Among these, DecompScore goes further by formalizing the decomposition task and introducing a metric for *atomicity*, or granularity, to compare decomposition methods. The measure counts the number of units that can be faithfully inferred from the source text, thereby favoring approaches that produce a larger set of accurate, fine-grained meaning units. **CORE** (Jiang et al., 2025) extends this by postprocessing decompositions to remove cross-unit redundancies, ensuring that the metric is not inflated by overlapping or paraphrastic facts. In this paper, we adopt the DecompScore granularity measure, combined with CORE's redundancy control, to assess QA-Noun together with QA-SRL, showing substantial gains in granularity compared to prior fact-based decomposition methods (§7).

While recent work advocates decomposing text into fact-like units, classical linguistically oriented representations — such as SRL, dependency-based semantics, AMR, and others — model textual meaning through structured predicate–argument relations (Abend and Rappoport, 2017). These frameworks are grounded in **Neo-Davidsonian semantics** (Parsons, 1990), where events and entities are represented as variables linked by binary relations labeled with thematic roles. For example, "The president signed the bill" is represented with an event `sign(e)`, entities `president(p)` and `bill(b)`, and relations `Agent(e,p)` and `Theme(e,b)`. QA-Noun builds on this tradition in a natural-language QA format, providing question–answer representations for nominal relations that, together with verbal QA-SRL, yield a structured, fine-grained decomposition of sentence meaning grounded in predicate–argument structure.

3 The QA-Noun Task

We introduce the QA-Noun task: a structured approach for identifying the semantic arguments of nouns in context. Given a sentence and a marked noun, our goal is to identify other phrases from the sentence that pertain to the noun, and represent its semantic relation to the noun using a simple question. For example, see Figure 1 for the arguments and their corresponding roles for the noun **album**.

In our task, we represent common semantic relations expressed by nouns through questions gener-



Question Type	Question Template
Location	Where is the [NOUN]?
Time	When is the [NOUN]?
Quantity	How much/ How many [NOUN]?
Partitive/ Membership (1)	What is the [NOUN] a part/member of?
Partitive/ Membership (2)	What/ Who is a part/member of [NOUN]?
Copular	What/ Who is (the) [NOUN]?
Property	What is the [PROPERTY] of [NOUN]?
Possession	Whose [NOUN]?
Sub-Specification	What kind of [NOUN]?

Figure 2: Example questions illustrating QA-Noun question templates. NOUN refers to the target noun.

ated from a carefully designed set of templates.

To ensure broad applicability, we place no restrictions on the types of nouns considered as predicates: any noun, regardless of its lexical category, or role in the sentence, is treated as a potential noun predicate.

3.1 Question Templates.

We define nine core question templates to capture the main argument types expressed by nouns (Figure 2). Examples include:

`Possessive` –	his *wife's* late **aunt**.
`Locative` –	the *Paris* **bridge**.
`Partitive` –	an *army* **officer**.

Our design follows the QASem tradition of systematically crafted question templates to capture predicate–argument relations (He et al., 2015; Pyatkin et al., 2020; Klein et al., 2020; Pesahov et al., 2023). The resulting templates provide interpretable and controllable mappings between linguistic structure and semantic role expression, making them well-suited for both annotation and model supervision.

Flexible Template Realization. While most templates are fixed, several allow minor modifications to better fit the sentence context and improve naturalness. For example, the `Partitive` template ***What/Who is a part/member of [NOUN]?*** permits annotators to choose between *part* or *member*, and between *what* or *who*, depending on the noun's semantics and discourse context. This controlled flexibility preserves the discrete, role-oriented character of the framework while enabling smoother phrasing and context-sensitive adaptation.

Hybrid Labeling via the Property Template. One template, `Property`, is reserved for open-ended attributes. It introduces a placeholder for a context-specific descriptor drawn from an open vocabulary, enabling flexible coverage of semantic properties beyond fixed roles. For example, in the sentence *"Valley Ranch is the team's 30-acre practice **camp**"*, the property template produces:
What is the [purpose] of the camp? → practice
What is the [size] of the camp? → 30-acre
Typical property values include *name*, *purpose*, *cause*, and *status*, inferred directly from context.

This hybrid design uniquely combines discrete, interpretable question types with an open-ended semantic slot. The result is a labeling scheme that offers both **consistency** (through fixed question types) and **expressivity** (via open-vocabulary properties), bridging the gap between structured role labeling and fully free-form natural language semantics (Michael et al., 2018).

Template Development and Validation. To ensure comprehensive coverage of noun-centered semantic relations, we drew on prior linguistic resources on nouns and noun compounds (Meyers et al., 2004; Tratz and Hovy, 2010). Starting from NomBank-aligned categories, we iteratively refined the templates through controlled annotation rounds. In early stages, we experimented with a larger set of candidate templates and crowdsourced dozens of sentences with multiple workers, analyzing the resulting **confusion matrix** of annotator choices to identify overlapping or ambiguous roles (e.g., between `Partitive` and `Membership`). We then abstracted and merged such categories to achieve a more discrete and distinguishable inventory, repeating this process until the final nine-template set reached stable coverage and low annotator confusion.

3.2 Argument and Question Scope

In QA-Noun, each semantic argument is represented as a contiguous phrase, that answers one of our template-based questions. The QA-Noun task is designed to complement verbal SRL by focusing specifically on noun-centered semantic relations that are not addressed through verb-based annotation. Thus we refrain from annotating arguments that would otherwise have been included in a semantic analysis of verbs in the sentence.

Annotators are instructed to select the most specific question template appropriate to the context, ensuring that the assigned role precisely captures the semantic relation of the argument to the noun. Nevertheless, multiple questions can often validly

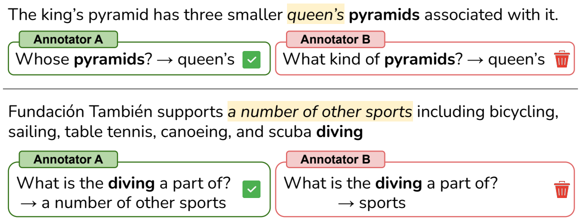

The king's pyramid has three smaller queen's pyramids associated with it.
Annotator A
Whose pyramids? → queen's
Annotator B
What kind of pyramids? → queen's
Fundación También supports a number of other sports including bicycling, sailing, table tennis, canoeing, and scuba diving
Annotator A
What is the diving a part of? → a number of other sports
Annotator B
What is the diving a part of? → sports

Figure 3: Reconciliation phase between two QA-Noun annotators. The annotators adjudicate between two proposed arguments that disagree either by extent or by semantic role. The selected argument and role after adjudication is shown schematically under annotator A (in green), while the discarded argument-role is shown under annotator B (in red). The top example showcases role (question) disagreement between the two annotators, while the bottom example depicts different extents (phrases) of the same argument.

apply to the same argument (e.g., *What is the location of X?* vs. *Where is X?*). We view this as an inherent feature of using natural language to represent semantics rather than a deficiency: each overlapping QA provides a complementary perspective on the relation. Downstream systems may aggregate such overlapping labels or exploit them as multi-faceted evidence for richer semantic modeling.

4 Dataset Construction

Data To create the QA-Noun dataset, we annotated over 1600 sentences with over 2000 noun mentions[1] across two main domains: Wikinews and Wikipedia. We annotated 1,686 sentences encompassing 2,029 nominal predicates, yielding a total of 4,869 arguments. The dataset was split into 50/10/40 (%) for the train, development and test splits. See Table 1 for template statistics.

Annotation Process We employed in-house annotators, primarily linguistics students or experienced English users (e.g., writers and language instructors). Following a controlled onboarding procedure inspired by Roit et al. (2020), candidates first underwent screening for English proficiency and fluency, then completed a paid training phase. During training, annotators studied detailed task guidelines[2], reviewed illustrative examples, and annotated a practice set of several dozen examples drawn from an *expert set* of 80 gold instances. Each candidate received personalized feedback based on automatic comparisons to gold annotations and follow-up meetings with the first author. This process ensured consistent, high-quality annotation before contributing to the main dataset.

During pilot studies, we observed, similarly to Roit et al. (2020), that argument identification from individual crowd workers often lacked sufficient coverage. To address this problem, we employ a two-annotator protocol to annotate a single predicate. First, two trained annotators are given a shared set of target predicates. Each annotator *independently* produces QA-pairs for each noun in the set according to our guidelines, using a dedicated annotation interface designed to streamline question formulation and answer span selection (see Appendix A.3 for details).

Consolidation After independent annotation, each pair of annotators meets online to reconcile differences. They review each other's argument spans and question templates, resolving discrepancies to reach a single agreed set. Missed arguments can be added, erroneous spans discarded, and questions refined to better capture the noun–argument relation. The final QA pairs are thus double-verified for accuracy while combining the coverage of two independent passes. Figure 3 illustrates this process: in one case, the more context-specific question is selected; in another, the more precise span is retained.

We employ this two-step protocol to collect our high-quality evaluation benchmark, while our training data is collected using a single trained annotator. This follows common practice in semantic annotation (e.g., Roit et al., 2020; FitzGerald et al., 2018; Kwiatkowski et al., 2019), where the training data undergoes lighter quality control to enable greater diversity and scale, while the evaluation sets are double-annotated and adjudicated to ensure reliability. Details regarding annotator compensation and cost breakdown are provided in Appendix A.4.

5 Assessing QA-Noun Dataset Quality

5.1 Evaluation Metrics

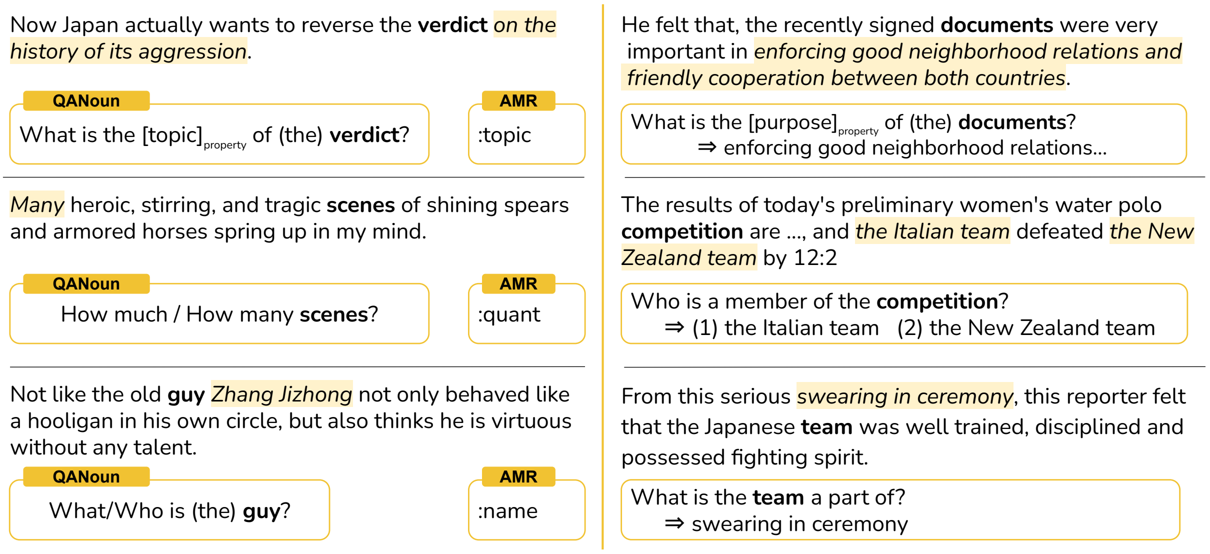
Similar to SRL, QA-Noun evaluation measures two abilities — correctly detecting the noun's arguments, and the correct assignment of semantic roles to these arguments. Following previous work (Pesahov et al., 2023; Roit et al., 2020; Pyatkin et al., 2020; Klein et al., 2020), we report standard precision and recall scores for unlabeled argu-

[1] In each sentence, we identify the target nouns using SpaCy's POS-tagger.

[2] The QANoun guidelines are publicly available at this slideshow.

	Property	Possession	Location	Quantity	Partitive/ Membership (1)	Partitive/ Membership (1)	Copular	Sub-Specification	Time
Total	1146	740	290	184	586	600	302	921	140

Table 1: Template statistics of the QA-Noun dataset



Now Japan actually wants to reverse the **verdict** *on the history of its aggression*.

QANoun: What is the $[\text{topic}]_{\text{property}}$ of (the) **verdict**?

AMR: :topic

Many heroic, stirring, and tragic **scenes** of shining spears and armored horses spring up in my mind.

QANoun: How much / How many **scenes**?

AMR: :quant

Not like the old **guy** *Zhang Jizhong* not only behaved like a hooligan in his own circle, but also thinks he is virtuous without any talent.

QANoun: What/Who is (the) **guy**?

AMR: :name

He felt that, the recently signed **documents** were very important in *enforcing good neighborhood relations and friendly cooperation between both countries*.

What is the $[\text{purpose}]_{\text{property}}$ of (the) **documents**?
⇒ enforcing good neighborhood relations...

The results of today's preliminary women's water polo **competition** are ..., and *the Italian team* defeated *the New Zealand team* by 12:2

Who is a member of the **competition**?
⇒ (1) the Italian team (2) the New Zealand team

From this serious *swearing in ceremony*, this reporter felt that the Japanese **team** was well trained, disciplined and possessed fighting spirit.

What is the **team** a part of?
⇒ swearing in ceremony

Figure 4: Comparison between example sentences with AMR and QA-Noun annotations. The noun predicate in each sentence is marked in bold and its argument is highlighted inline. **Left** Comparison between semantic roles when the argument is mutually annotated. **Right** Diverse arguments captured by QA-Noun's annotators that were out of scope for AMR. They represent different implied meanings, memberships and other relations.

ment detection (UA) against a set of ground-truth arguments. Briefly, a predicted argument is considered to be correct if it significantly overlaps with a gold argument, with a token-level intersection over union greater than 0.5. We apply maximal bipartite matching to enforce a one-to-one alignment between predicted and gold arguments, weighting each pair by their overlap score. We count the number of true positives as the number of matches, and false positives and false negatives as the number of leftover arguments from the predicted and ground truth sets, respectively.

In contrast to argument detection, evaluating the accuracy of semantic role assignment is a challenge. In QA-Noun, the roles are represented as question templates, and they are not mutually exclusive. For instance, as shown in Figure 1, the argument *1971* could be annotated both with *What is the [year] of the album?* and *When is the album?*. Therefore, comparing against a single ground-truth template could underestimate role assignment accuracy.

To address this, we manually evaluate whether the selected question template accurately captures the semantic relationship between the noun and the predicted answer span. This evaluation is performed on correctly predicted arguments, and we report the proportion of *sound role assignments* (SRA). To further increase reliability, two experts independently assess each role assignment, and report the average SRA from their evaluations.

5.2 Comparison with AMR

We compared QA-Noun annotations and AMR structures over the same set of nouns to gain insight about their relative coverage in the scope of annotation. Given a predicate noun, we manually align its QA-Noun arguments with the associated AMR entities and analyze the differences. Since AMR represents the sentence using a directed graph over *entities*, while QA-Noun annotates *lexical units* in the sentence, we first identify the node in the AMR graph that represents the predicate entity and extract its arguments.[3] In particular, we consider both directions to and from the predicate node in the AMR graph as plausible arguments.

In this analysis, we annotated a sample of 40 nouns from the AMR Bank, yielding 156 QA-Noun arguments, while the corresponding AMR entities include 90 arguments in total. Our analysis showed that QA-Noun captures almost all noun-related relations represented in AMR (**89/90; recall = 0.99,**

[3] At times, the noun's direct parent in AMR refers to the same entity, and in that case we take the parent's arguments as well.

95% CI [0.97, 1.00], bootstrap 200K replicates), with most question templates aligning closely to AMR roles (Figure 4). Beyond this significant overlap, QA-Noun has annotated 65 additional arguments absent from AMR, including implied relations (21), membership roles (16), coreferent mentions (13) and various other cases, alongside 4 annotation errors. While some of the implied relations are inferential and out of scope for AMR, some are captured implicitly in AMR's deeper graph structure but are not linked as direct arguments.

These promising results, both in correctness of the QA-Noun arguments and the almost full coverage of AMR entities, suggest that QA-Noun contributes reliably annotated semantic relations that AMR often leaves implicit or underspecified. QA-Noun thus provides a broader and more accessible representation of noun arguments, including co-referential mentions and implied roles that are difficult to recover from AMR alone. Its natural-language question format enables detailed semantic coverage while remaining intuitive and scalable for annotation, making it a strong complement to existing structured frameworks akin to AMR.

5.3 Inter-Annotator Agreement

To estimate the consistency of the dataset across different annotations, we measure inter-annotator agreement (IAA) on a sample of 90 target nouns. While worker-vs-worker agreement for QAs is somewhat partial — mostly due to insufficient coverage, as discussed above (§4) — the overall consistency of the dataset is assessed by comparing consolidated annotations obtained from disjoint pairs of workers after adjudication. The macro-averaged unlabeled agreement (UA) F1 score for inter-annotator agreement is **72.8**.

Although somewhat lower than the expert agreement levels typically reported for tightly constrained schemes such as NomBank, our IAA reflects the open-ended nature of QA-Noun's semantic coverage and the expressivity of the QA format, and is comparable to agreement levels reported for other QASem tasks (Klein et al., 2020; Pesahov et al., 2023).

6 Modeling

In addition to defining the QA-Noun representation and dataset, our goal is to develop and release a practical tool for semantic decomposition — one that is both accurate and efficient. To this end, we experiment with two modeling approaches: in-context prompting and parameter-efficient fine-tuning, evaluating their performance on the QA-Noun task. Data, models and experiments code can be found in the project repository.[4]

	Model		Precision	Recall	F1
ICL	Llama 3	8B	56.4	35.5	43.6
	Llama 3	70B	67.4	40.2	50.4
	Llama 3.1	405B	**64.7**	51.0	**57.0**
FT	Llama 3	8B	49.7	**62.7**	55.4
	Qwen 2.5	14B	62.5	48.1	54.4
	Phi 4	14B	49.1	57.5	53.0

Table 2: Different models automatic evaluation results against our test set. All reported metrics are for unlabeled argument detection (UA). ICL stands for In-Context Learning methods, while FT is for fine-tuned methods.

6.1 Methods

In-Context Learning (ICL). We evaluate several large language models (LLMs) using few-shot prompting. Each model is prompted to generate all relevant QA pairs for a target noun in context using our predefined question templates, with at least two examples per template included in the prompt to guide completions. The full prompt is provided in Appendix A.1. We test the following LLMs: `LLaMA-3-70B`, `LLaMA-3-8B` and `LLaMA-3.1-405B`.

LoRA Fine-Tuning (FT). To adapt moderately sized models for the task, we apply Low-Rank Adaptation (LoRA; Hu et al., 2022), updating only a small subset of parameters during training. We fine-tune three models — `LLaMA-3-8B`, `Qwen-2.5-14B`, and `Phi-4-14B` — on our QA-Noun training set, using gold question-answer pairs as supervision. Multiple hyperparameter configurations were explored to optimize performance (see Appendix A.2 for details).

6.2 Model Evaluation

Main Results. Table 2 presents evaluation results on our test set, comparing in-context and fine-tuned models in terms of unlabeled argument detection. The best overall performance is achieved by `LLaMA-3.1-405B` in the in-context setting, demonstrating strong generalization even without task-specific training. Notably, the fine-tuned `LLaMA-3-8B` performs competitively, outper-

[4] https://github.com/unimaria/QA-Noun

forming all other fine-tuned and in-context models apart from `LLaMA-3.1-405B`.

These results highlight the value of task-specific supervision: with a modest parameter footprint, fine-tuned models approach the performance of much larger LLMs — making them practical for large-scale decomposition pipelines where efficiency is essential. For this reason, we select the fine-tuned `LLaMA-3-8B` as our parser backbone, balancing strong performance with open licensing and cost-efficiency for scalable downstream use.

Role Assignment As discussed in Section 5.1, the flexible nature of QA-Noun question templates poses a challenge for automatic evaluation against a ground truth that contains only a single label. To address this, we conducted both manual and automatic evaluations of role assignment for our selected model, a fine-tuned `LLaMA-3-8B`.

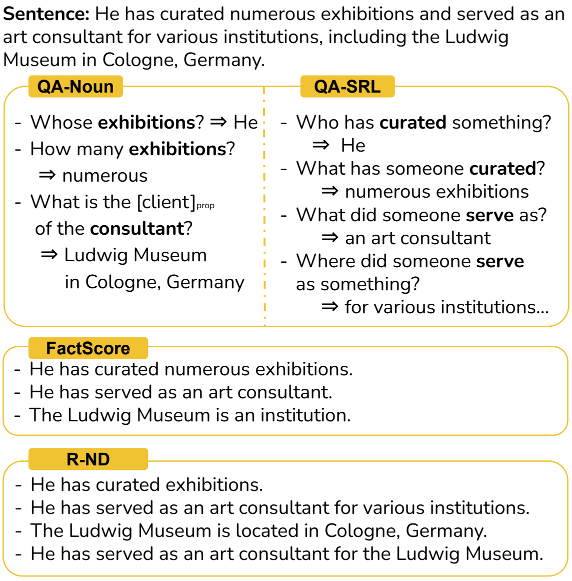
In the manual evaluation, two of the authors independently reviewed model-generated questions for correctly identified arguments across 54 target nouns (115 QA pairs in total). The resulting average semantic-role accuracy (SRA) was **58.5%**, indicating that while arguments are generally recovered reliably, the model often struggles to select the most contextually appropriate question template.

To complement this small-scale analysis, we performed an automatic evaluation using a strong LLM (`GPT-4o`) as an entailment-based judge. For 1,425 QA pairs where the predicted argument span was correct, the model was asked whether each QA pair was entailed by the original sentence — interpreted as a proxy for question validity. Approximately **65%** of the QAs were judged valid, a slightly higher rate than the manual estimate. Together, these analyses suggest that while QA-Noun effectively captures most nominal relations, selecting the most fine-grained and semantically precise role remains a key challenge, motivating future work on improved modeling and richer supervision.

We next move from argument-level accuracy to evaluating QA-Noun+QA-SRL as a decomposition framework, comparing its granularity to fact-based methods such as FactScore and DecompScore.

7 Granular Information Decomposition

As discussed in the introduction, capturing sentence meaning via a maximally atomic decomposition of information units is key for modeling semantic alignment. QA-Noun combined with QA-SRL produces explicit predicate–argument QA pairs as granular meaning units, in contrast to recent "atomic fact" approaches that prompt LLMs to generate unconstrained declarative statements. To quantify this difference, we evaluate the granularity of our QA-based decompositions using **DecompScore** (Wanner et al., 2024), a metric designed to assess the granularity and coverage of decomposition methods, as described in Section 2.3. We adapt it to treat each QA pair as an atomic sub-claim and compare against the decompositions of FactScore



Sentence: He has curated numerous exhibitions and served as an art consultant for various institutions, including the Ludwig Museum in Cologne, Germany.

QA-Noun
- Whose **exhibitions**? ⇒ He
- How many **exhibitions**? ⇒ numerous
- What is the [client]$_{prop}$ of the **consultant**? ⇒ Ludwig Museum in Cologne, Germany

QA-SRL
- Who has **curated** something? ⇒ He
- What has someone **curated**? ⇒ numerous exhibitions
- What did someone **serve** as? ⇒ an art consultant
- Where did someone **serve** as something? ⇒ for various institutions...

FactScore
- He has curated numerous exhibitions.
- He has served as an art consultant.
- The Ludwig Museum is an institution.

R-ND
- He has curated exhibitions.
- He has served as an art consultant for various institutions.
- The Ludwig Museum is located in Cologne, Germany.
- He has served as an art consultant for the Ludwig Museum.

Figure 5: Sentence decomposition with QA-Noun and QA-SRL compared to fact-based approaches. QA-Noun captures noun-centered relations (e.g., *Whose exhibitions?*) and surfaces implicit links such as client–consultant relations, which fall under inferential arguments. Combined with QA-SRL verbal roles, they yield a structured predicate–argument breakdown. FactScore and R-ND generate declarative "atomic facts" but typically do not capture such inferential relations.

DecompScore: Entailed Sub-claims Per Sentence

Method	Generated	Non-Redundant	Entailed
FactScore (GPT-4o)	4.9	3.2	3.1 ± 0.1
R-ND (GPT-4o)	5.4	3.7	3.7 ± 0.1
QASem (Llama-3-8B)	14.1	7.2	**4.8** ± 0.2

Table 3: DecompScore's decomposition *atomicity* metric: The number of sub-claims/QAs generated by each method per sentence, broken down by the number of generated, non-redundant, and finally entailed units. The *Entailed* column is the final DecompScore metric, calculated with a 95% confidence interval. Our approach (QA-Noun + QA-SRL) is denoted as QASem.

(Min et al., 2023) and R-ND (DecompScore's own GPT-based method).

Setup. We randomly sample 1,000 sentences from the FactScore benchmark, following the evaluation protocol of Wanner et al. (2024) for comparability. For each sentence, we generate question–answer pairs using our fine-tuned `LLaMA-3-8B` QA-Noun parser and the QA-SRL parser (Klein et al., 2022). Each QA pair is treated as a candidate meaning unit and evaluated for total count, non-redundant count after CORE filtering, and source entailment via DecompScore's GPT-based pipeline.

Because the verbal and nominal parsers target different parts of speech, they occasionally recover overlapping semantic relations. To avoid double-counting the same sub-claim, we use the **CORE** framework (Jiang et al., 2025) to automatically identify and cluster redundant or paraphrastic QA pairs. This ensures a faithful count of unique atomic facts while still allowing cross-validation of equivalent relations across syntactic forms. Notably, these overlaps reflect distinct yet complementary linguistic realizations of the same content, and may offer added value in enriching semantic labels for certain downstream tasks. See Appendix A.5 for examples and discussion of common overlap patterns.

Notably, unlike FactScore and R-ND GPT-4o–based baselines, our entire pipeline uses open-weight models to enable reproducible, scalable deployment.

Results. Table 3 summarizes the results. The combined QA-Noun+QA-SRL system yields over **150%** more validated, non-redundant semantic units than FactScore and about **130%** more than DecompScore, highlighting the granularity advantage of a structured predicate–argument approach over unconstrained fact extraction. Nominal QA pairs account for nearly half of the decomposition output (2.4 QA pairs per sentence on average vs. 2.3 from verbs), underscoring the necessity of modeling noun-centered semantics for complete sentence-level meaning and cross-text alignment.

Although GPT-4o baselines achieve slightly higher entailment precision, our approach delivers substantially greater coverage and atomicity. This reflects an inherent trade-off: exhaustively recovering predicate–argument structures for both verbs and nouns is a harder task than producing unconstrained facts, yet it yields decompositions that are more linguistically grounded and robust. Importantly, many of the redundant QAs filtered by CORE are *valid paraphrases* rather than noise, providing complementary surface realizations of the same fact — an aspect future systems could exploit for richer semantic labeling or multi-view learning.

An illustration of the resulting decomposition is shown in Figure 5. Taken together, these findings position QA-Noun+QA-SRL — and by extension the broader QASem framework — as a comprehensive and structured alternative to fact-based decomposition methods for applications requiring precise cross-text semantic alignment.

8 Conclusion

We introduced QA-Noun, a framework for representing noun semantics as a set of QAs, each expressing a predicate-argument level fact involving the noun, which integrates seamlessly into the broader QA-based semantic paradigm. By combining templated and open-ended questions, QA-Noun provides a scalable and interpretable method for capturing noun-centered meaning in context. Our dataset and evaluations demonstrate its reliability and ability to capture rich, fine-grained relations, making it a strong foundation for highly-granular decomposition of textual meaning, as needed for cross-text semantic alignment. Together with QA-SRL and related QA-based tasks, QA-Noun advances toward a comprehensive framework for sentence-level semantic decomposition.

Limitations

While QA-Noun provides a structured and scalable approach for semantic role labeling of noun predicates, several limitations remain in its current form.

Training Data Quality. As discussed in Section 4, our training data was single-annotated to enable greater diversity and scale within the available budget, while double annotation and adjudication were reserved for the evaluation sets. This practical design choice follows established practice but inevitably introduces some annotation noise and variability, particularly for subtle or context-dependent roles. Such inconsistencies may limit the ultimate performance of models trained solely on the training data, motivating future work on selective re-annotation or semi-automatic verification of difficult cases.

Question Template Design. The fixed set of question templates enables consistency and interpretability, but also imposes important limitations. First, many of the templates are not phrased in fully natural language, which may hinder large language models (LLMs) that rely on surface form likelihoods for generation. This can lead to errors when models favor more common but less semantically appropriate templates. Second, the templates are not strictly mutually exclusive — different questions can validly apply to the same argument span — posing challenges during both training and evaluation. Disambiguating between overlapping templates remains a non-trivial problem for both humans and models.

Model Efficiency and Scale. Our best-performing model in terms of raw F1 is a large in-context LLM (LLaMA 405B), which is expensive to run and unsuitable for deployment at scale. While our fine-tuned models (e.g., LLaMA 8B) offer a more practical solution, they still fall short of state-of-the-art LLMs in some scenarios.

Domain Generalization. QA-Noun is built from and evaluated on formal text from Wikipedia and Wikinews. Although diverse, these sources do not capture the full spectrum of language use. It is unclear how well models trained on QA-Noun would generalize to other genres such as narrative, conversational, or low-resource domains.

References

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Daniela Brook Weiss, Paul Roit, Ayal Klein, Ori Ernst, and Ido Dagan. 2021. QA-align: Representing cross-text content overlap by aligning question-answer propositions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9879–9894, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arie Cattan, Paul Roit, Shiyue Zhang, David Wan, Roee Aharoni, Idan Szpektor, Mohit Bansal, and Ido Dagan. 2024. Localizing factual inconsistencies in attributable text generation.

Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth, and Tal Schuster. 2023. Propsegment: A large-scale corpus for proposition-level segmentation and entailment recognition.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Yanai Elazar, Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2022. Text-based NP enrichment. *Transactions of the Association for Computational Linguistics*, 10:764–784.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *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)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Jing Fan, Dennis Aumiller, and Michael Gertz. 2023. Evaluating factual consistency of texts with semantic role labeling. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 89–100, Toronto, Canada. Association for Computational Linguistics.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060, Melbourne, Australia. Association for Computational Linguistics.

Mantas Gavenavicius. 2020. Evaluating and comparing textual summaries using question answering models and reading comprehension datasets. B.S. thesis, University of Twente.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2:: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zheng Ping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi Liu, and Benjamin Van Durme. 2025. Core: Robust factual precision with informative sub-claim identification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19833–19856.

Ayal Klein, Eran Hirsch, Ron Eliav, Valentina Pyatkin, Avi Caciularu, and Ido Dagan. 2022. QASem parsing: Text-to-text modeling of QA-based semantics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7742–7756, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. Annotating noun argument structure for NomBank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke S. Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *NAACL-HLT*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. *arXiv*, abs/2305.14251.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint.

Roberto Navigli, Marco Pinto, Pasquale Silvestri, Dennis Rotondi, Simone Ciciliano, and Alessandro Scirè. 2024. Nounatlas: Filling the gap in nominal semantic role labeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16245–16258.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Terence Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press, Cambridge, MA.

Leon Pesahov, Ayal Klein, and Ido Dagan. 2023. QA-adj: Adding adjectives to QA-based semantics. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 74–88, Nancy, France. Association for Computational Linguistics.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Haoyi Qiu, Kung-Hsiang Huang, Jingnong Qu, and Nanyun Peng. 2024. AMRFact: Enhancing summarization factuality evaluation with AMR-driven negative samples generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 594–608, Mexico City, Mexico. Association for Computational Linguistics.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Paul Roit, Aviv Slobodkin, Eran Hirsch, Arie Cattan, Ayal Klein, Valentina Pyatkin, and Ido Dagan. 2024. Explicating the implicit: Argument detection beyond sentence boundaries. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16394–16409, Bangkok, Thailand. Association for Computational Linguistics.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Uppsala, Sweden. Association for Computational Linguistics.

Shayne Longpre Wanner, Jason Weston, Mikel Artetxe, Arthur Szlam, and Angela Fan. 2024. DecompScore: Reference-based evaluation of sentence decompositions into atomic facts. In *arXiv preprint*, arXiv. arXiv. ArXiv:2403.11903.

Shiyue Zhang, David Wan, Arie Cattan, Ayal Klein, Ido Dagan, and Mohit Bansal. 2025. QAPyramid: Fine-grained Evaluation of Content Selection for Text Summarization. In *Proceedings of the Conference on Language Modeling (COLM)*. Association for Computational Linguistics.

Fangwei Zhu, Peiyi Wang, and Zhifang Sui. 2024. Reducing hallucinations in entity abstract summarization with facts-template decomposition.

A Appendix

A.1 Model

The prompt used for the in-context and fine-tuning experiments is shown in Table 4.

A.2 Fine-Tuning Configurations

We performed a targeted hyperparameter search on the LLaMA 8B (v3) model, experimenting with a range of LoRA settings that varied in rank, scaling factor (α), and training epochs (see Table 5). All experiments used a learning rate of 0.0002 and the AdamW optimizer. For the larger models, Qwen 14B (v2.5) and Phi 14B (v4), we did not perform a full hyperparameter sweep due to compute constraints. Instead, we applied promising configurations observed during LLaMA tuning. For Qwen, we reused the best-performing LLaMA settings (rank 64, α 16), while for Phi we adopted a configuration (rank 32, α 8) that was reported to perform well in public benchmarks.

Among all configurations, the strongest results on the QA-Noun development set were achieved by:

`LlaMA-3-8B` with rank 64, $\alpha = 16$

`LlaMA-3-8B` with rank 8, $\alpha = 32$

`Phi-4-14B` with rank 32, $\alpha = 8$

The best-performing configuration on the QA-Noun development set was `LlaMA-3-8B` with rank 64 and $\alpha = 16$, which we selected for our parser.

A.3 Annotation Interface

We developed a dedicated Graphical User Interface (GUI) (see Figure 6) that presented annotators with a list of sentences, each containing a highlighted noun, and tasked them with generating question-answer pairs specific to the noun. To create the questions, annotators first selected an appropriate question template from a drop-down menu and populated any required slots for the selected template. Then, they marked the corresponding argument-answer by selecting a contiguous span of text within the sentence.

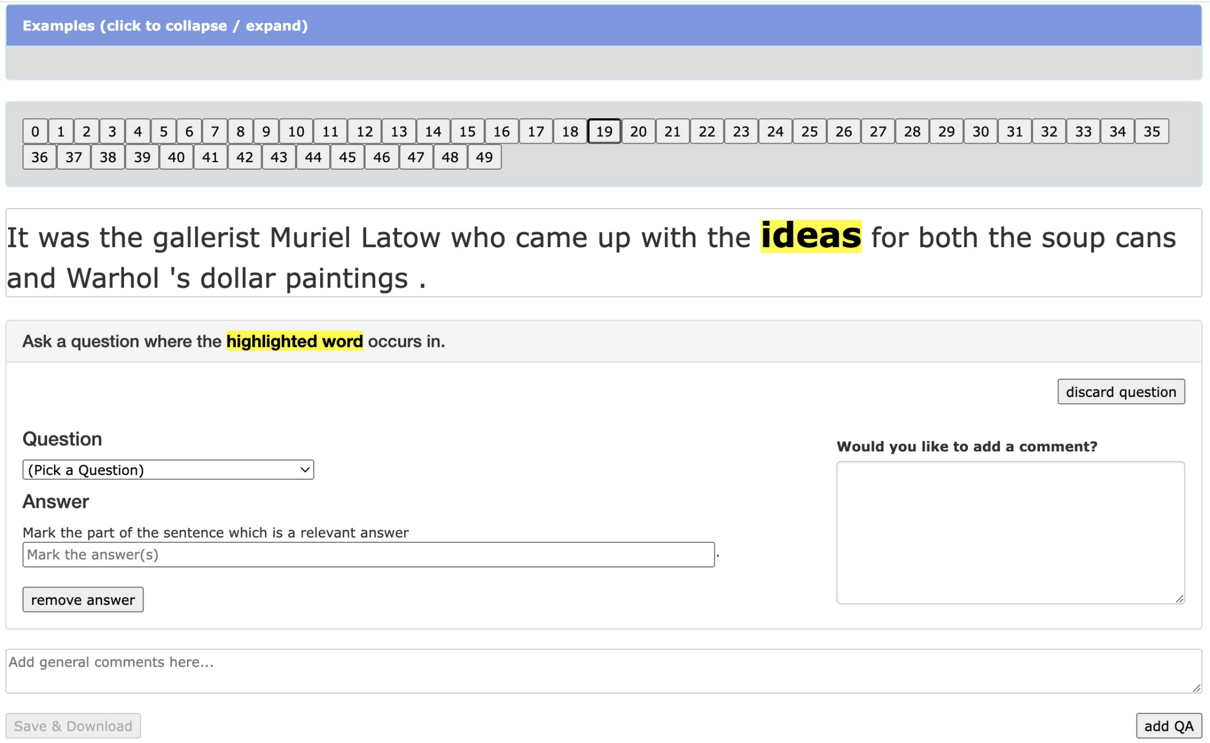
A.4 Annotation Costs

Our annotators were paid $13 per hour for both generation and reconciliation steps, which is approximately 170% of the local minimum wage. This resulted in an average cost of $1.75 per predicate in the evaluation set — compensating for two annotation steps and a reconciliation session. For the training set, which employed a single annotator step, the average cost per predicate was $0.35. In total, the cost of dataset curation and annotator onboarding is estimated at approximately $2,980.

A.5 Overlap Between QA-Noun and QA-SRL

Although QA-Noun and QA-SRL annotate complementary syntactic structures, they often recover semantically equivalent relations. To ensure accurate counts of unique content units during evaluation, we apply the **CORE** framework to detect and remove such redundancies.

In Table 6 we present a range of overlap examples filtered out by CORE and categorize them into common types of semantic overlap. We find that **Agent** overlaps — e.g., a noun's possessor versus the subject of a verb — are the most common, reflecting shared underlying predicate-argument structures. **Location** overlaps are also frequent when both noun phrases and verb predicates reference spatial context. **Time** overlaps arise when temporal markers are linked to both event-denoting nouns and corresponding verbal mentions. Other overlap types include **Purpose**, **Membership**, and **Possession**. In all cases, CORE helps ensure that overlapping facts are scored only once while preserving their alignment for downstream applications such as paraphrase learning or redundancy detection.



It was the gallerist Muriel Latow who came up with the **ideas** for both the soup cans and Warhol 's dollar paintings .

Ask a question where the **highlighted word** occurs in.

Mark the part of the sentence which is a relevant answer

Figure 6: Graphical User Interface (GUI) for annotators.

Read the Sentence and focus on the noun that is marked with <f></f>.
Find all the words or short phrases which provide information about the noun entity of the noun marked in <f></f>- they are called arguments.
The arguments should be a continuous span from the sentence, they should appear with the exact words and order as appeared in the sentence.
Use the arguments you found as answers, and generate questions to match those answers.
The questions should be taken from the list of templates:
1: What is the <property> of (the) <f>noun</f>?,
2: Whose <f>noun</f>?,
3: Where is the <f>noun</f>?,
4: How much /How many <f>noun</f>?,
5: What is the <f>noun</f> a part/member of?,
6: What/Who is a part/member of <f>noun</f>?,
7: What/Who is (the) <f>noun</f>?,
8: What kind of <f>noun</f>?,
9: When is the <f>noun</f>?
The number marks the template's number.
The <f>noun</f> should be replaced with the noun that is marked with <f></f> in the sentence.
The <property>tag should be replaced with a word that describes a property of the noun (color, size, cause etc.), that matches the answer.
Don't generate the same answer for two different questions, choose the most suitable question for each answer.
Display the list of QAs sorted in ascending order by question template id.
If you can't find any arguments to the noun marked in <f></f>, the output should be: "There are no QAs generated."

The format should be:
QAs:
Question template number: <the number>
Question: <the question>
Answer: <the answer>

Table 4: Prompt used for in-context and fine tuning tasks

Model	LoRA Rank	α	Epochs
LLaMA 8B (v3)	64	16	6
LLaMA 8B (v3)	32	128	3
LLaMA 8B (v3)	64	64	3
LLaMA 8B (v3)	16	64	10
LLaMA 8B (v3)	64	16	3
LLaMA 8B (v3)	16	32	20
LLaMA 8B (v3)	8	32	3
Qwen 14B (v2.5)	64	16	3
Phi 14B (v4)	32	8	3

Table 5: LoRA fine-tuning hyperparameter configurations explored for each model.

Sentence and QA Pairs	**Category**
Sentence: She has written articles and essays for various journals, edited volumes, and exhibition catalogs. QA-Noun: Whose **articles**? → *She* QA-SRL: Who has **written** something? → *She*	Agent
Sentence: Father Tompkins also played a significant role in shaping the labor movement in Nova Scotia. QA-Noun: Whose **role**? → *Father Tompkins* QA-SRL: Who **played** something? → *Father Tompkins* QA-Noun: Where is the **movement**? → *Nova Scotia* QA-SRL: Where did someone **play** something? → *Nova Scotia*	Agent, Location
Sentence: She served as Chair of the Department of Performing and Fine Arts from 2012 to 2019. QA-Noun: When is the **position**? → *2012–2019* QA-SRL: When did someone **serve** as something? → *2012–2019*	Time
Sentence: Over the course of his career, Nieves played for several MLB teams including the New York Yankees and others. QA-Noun: Whose **teams**? → *MLB (Nieves)* QA-SRL: What included something? → *several MLB teams*	Possession
Sentence: Tugman has utilized his skills to secure multiple victories on the professional circuit. QA-Noun: What is the **purpose** of the skills? → *to secure victories* QA-SRL: Why has someone **utilized** something? → *to secure victories*	Purpose

Table 6: Examples of overlapping semantic relations recovered by both QA-SRL (verbal) and QA-Noun (nominal) parsers and filtered by the CORE framework. Categories include Agent, Location, Time, and others such as Purpose and Possession.