

Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials: Methods

ISSN: 2766-0400 (Online) Journal homepage: www.tandfonline.com/journals/tstm20

Extraction of polymer biodegradation events from scientific papers: structured dataset design and contradiction-aware scientific event assessment

Shanshan Liu, Yuji Matsumoto & Masashi Ishii

To cite this article: Shanshan Liu, Yuji Matsumoto & Masashi Ishii (2026) Extraction of polymer biodegradation events from scientific papers: structured dataset design and contradiction-aware scientific event assessment, Science and Technology of Advanced Materials: Methods, 6:1, 2745582, DOI: 10.1080/27660400.2026.2745582

To link to this article: https://doi.org/10.1080/27660400.2026.2745582

© 2026 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group

Published online: 04 Nov 2026.

Submit your article to this journal

Article views: 46

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tstm20

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2026, VOL. 6, NO. 1, 2745582
https://doi.org/10.1080/27660400.2026.2745582

Taylor & Francis
by informa

OPEN ACCESS Check for updates

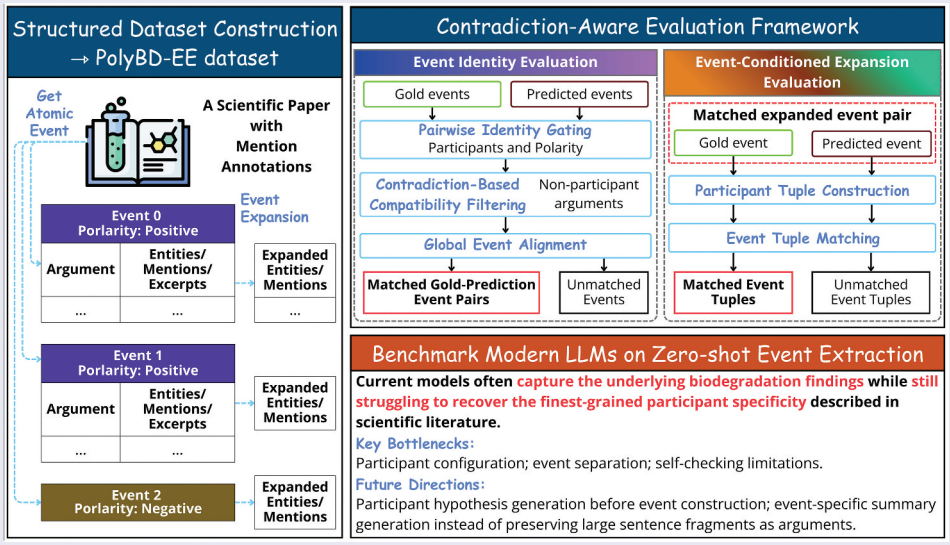
Extraction of polymer biodegradation events from scientific papers: structured dataset design and contradiction-aware scientific event assessment

Shanshan Liu [a,b], Yuji Matsumoto [c,d] and Masashi Ishii [a]

[a]Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Tsukuba, Japan; [b]Graduate School of Science and Technology, University of Tsukuba, Tsukuba, Japan; [c]Center for Language AI Research, Tohoku University, Sendai, Japan; [d]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

ABSTRACT

Scientific literature on polymer biodegradation contains experimentally important information regarding degradable polymers, microorganisms, environmental conditions, and degradation outcomes. However, biodegradation findings are often condition-sensitive and distributed across full-text articles, making reliable event extraction and evaluation difficult. In this work, we introduce PolyBD-EE, a document-level polymer biodegradation event extraction dataset with structured event annotations and text-grounded participant expansion. We further propose a contradiction-aware evaluation framework that evaluates whether predicted and gold events describe experimentally compatible biodegradation findings under different participant specificity levels. Using PolyBD-EE, we benchmark multiple modern generative large language models under zero-shot extraction settings. The best-performing setting, GPT-5.5-*Think*, achieves strict- and lenient-atomic F1 scores of 0.492 and 0.557, respectively, indicating that current models can capture underlying biodegradation findings but still struggle with fine-grained participant specificity described in the literature. Further analysis identifies key bottlenecks, including incorrect participant configurations, over-generalization during participant expansion, and weak separation of event-specific information. Manual validation of GPT-5.5-*Think* predictions confirms 96.9% and 96.6% of accepted alignments under strict and lenient evaluation, respectively, with false-rejection rates of 0.55% and 1.85%. These findings show that document-level biodegradation event extraction remains a challenging task with substantial room for improvement. PolyBD-EE and the proposed evaluation framework provide a task-specific basis for measuring future progress, diagnosing failure modes, and guiding the development of more reliable scientific event extraction systems for polymer biodegradation research and experimental materials science. Code and publicly redistributable data are available at: https://github.com/sl-633/polybd-ee-caef



IMPACT STATEMENT

We introduce PolyBD-EE dataset and a contradiction-aware evaluation framework, establishing a benchmark for document-level polymer biodegradation event extraction and supporting more reliable scientific information extraction in materials science.

ARTICLE HISTORY

Received 16 June 2026
Revised 31 August 2026
Accepted 4 October 2026

KEYWORDS

Event extraction; information extraction; large language model; polymer biodegradation; materials informatics

CONTACT Shanshan Liu liu.shanshan@nims.go.jp Center for Basic Research on Materials, National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

© 2026 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

1. Introduction

Scientific literature contains a large number of experimentally grounded findings describing how objects behave, interact, transform, or respond under specific experimental conditions. From an information extraction perspective, these experimentally grounded findings can be viewed as experimental scientific events involving participants, experimental conditions, and empirical outcomes.

Polymer biodegradation research provides a representative example of such events, where experimentally important evidence is often distributed across full-text articles and strongly influenced by material properties, environmental conditions, and microorganism specificity. Structuring such information automatically helps researchers organize biodegradation evidence across the literature, compare experimentally reported degradation behaviors, and accelerate the discovery of potentially degradable materials and microorganisms.

However, reliably structuring and evaluating such experimentally grounded scientific events remains challenging. Scientifically incompatible findings may still exhibit high textual or semantic similarity, while experimentally equivalent findings may be expressed through heterogeneous descriptions. For example, the statements '*UV-pretreated PE is degraded*' and '*non-pretreated PE is degraded*' exhibit high similarity under overlap-based measures such as Jaccard similarity as well as embedding-based similarity methods using MatBERT [1] representations, while describing experimentally incompatible biodegradation findings.

Existing event extraction frameworks are not well-suited for such experimentally grounded scientific events [2–8]. Most existing scientific event extraction frameworks remain trigger-centered and rely on local span-level representations for both annotation and evaluation. However, experimentally important information often extends beyond isolated trigger expressions or entity spans. As shown in Figure 1, statements such as '*The compost contained 5 wt% of the polyethylene on dry basis*' describe critical experimental conditions, while statements such as '*The smooth surface of the LMWPE sheets became eroded*' provide observed degradation evidence. Such information is difficult to represent naturally within conventional span-based extraction frameworks.

To address these challenges, we study polymer biodegradation event extraction as a representative experimental scientific event extraction task. We introduce PolyBD-EE, a document-level dataset for polymer biodegradation event extraction and evaluation. The dataset supports both biodegradation findings at different participant specificity levels and constrained scientific generalization across related polymers and microorganisms.

We further propose a contradiction-aware evaluation framework motivated by observations from preliminary human event-ranking analysis. Rather than evaluating predictions solely through span overlap or semantic similarity, the proposed framework evaluates whether predicted and gold events describe experimentally compatible biodegradation findings based on participant consistency, experimental conditions, and event granularity. The framework additionally supports the evaluation of event expansion by considering participants at different levels of specificity (e.g. low-molecular-weight polyethylene vs. polyethylene, or microbial strains vs. species/genus-level representations).

Finally, using PolyBD-EE, we benchmark multiple modern generative large language models under multi-turn zero-shot extraction settings. Experimental results reveal performance gaps between strict and lenient event identity evaluation, suggesting that current models can capture the underlying biodegradation finding while still struggling to recover the finest-grained participant specificity described in the literature. We also observe that reasoning-based generation improves event identification for several models, although the gains are not consistent across models. Participant expansion remains challenging, with models frequently over-generalizing to broader participant concepts.

Our analysis further reveals two major bottlenecks in document-level scientific event extraction. First, incorrect participant configurations remain a primary source of false positives, suggesting that future systems may benefit from mechanisms that improve participant-level event hypothesis generation before detailed event construction. Second, models often retain sentence-level evidence rather than extracting event-specific information, resulting in weak separation between event arguments and between atomic events. This suggests that constructing cleaner event representations may require event-specific evidence summarization mechanisms rather than directly extracting textual fragments from the source document.

Our main contributions are summarized as follows:

- We introduce PolyBD-EE, a document-level dataset for polymer biodegradation event extraction and evaluation, featuring structured event annotations, document-level entity linking, and text-grounded event expansion across scientifically related participant concepts.
- We propose a contradiction-aware evaluation framework for polymer biodegradation events, including strict and lenient event identity evaluation together with constrained event expansion evaluation under different participant specificity levels.

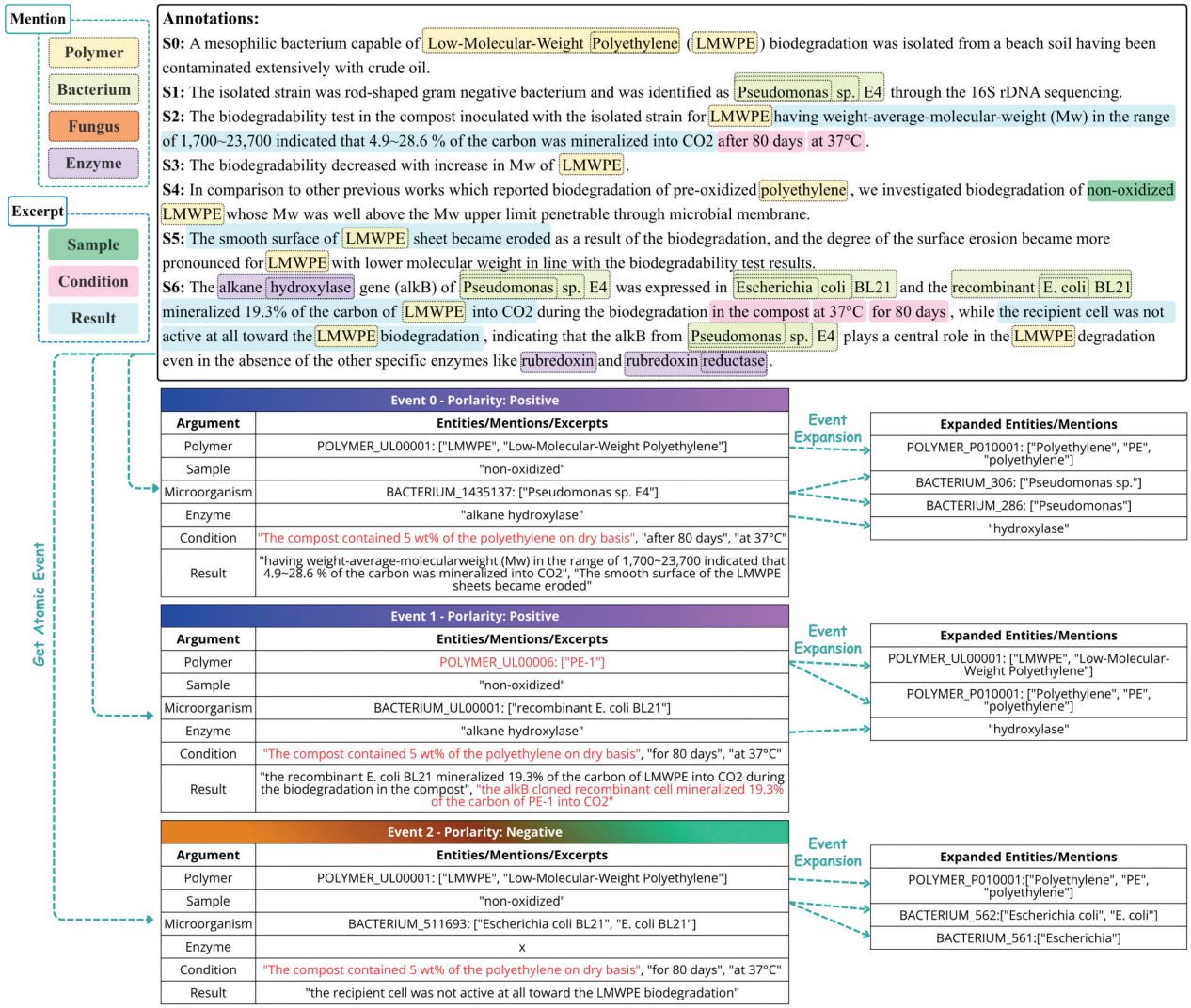


Figure 1. An example of entity and event annotations in PolyBD-EE based on text excerpts from a paper on polymer biodegradation [9]. Given the first seven sentences (*S0-S6*), three atomic events are annotated. Event information outside these sentences in the full paper is highlighted in red. Mentions of polymer, bacterium, and fungus are grouped into entities for event annotation.

- We conduct systematic zero-shot evaluations of multiple modern generative large language models under both instruct- and thinking-mode generation settings.
- Through the proposed evaluation framework, we identify key bottlenecks in current document-level biodegradation event extraction systems. We further suggest promising directions for future scientific event extraction systems based on these observations.

2. Background

2.1. Event extraction in materials science

Plastic waste accumulation in the environment has become a growing concern, motivating increasing interest in biodegradable plastics as potential alternatives to conventional plastics [10–14]. Previous reviews have noted that systematic analysis of biodegradation studies is complicated by substantial variations in materials, testing methods, and evaluation criteria [15]. Structured extraction of biodegradation findings from scientific literature can facilitate systematic analysis by organizing experimentally relevant information into structured representations. From an information extraction perspective, these requirements can be formulated as an event extraction problem.

Event extraction is a traditional information extraction task that aims to identify structured event representations from text. A widely adopted formulation originates from the ACE2005 corpus [2], where events are organized around trigger expressions and their associated arguments. Since ACE2005 was originally designed for general-domain text such as news articles, many domain-specific adaptations have subsequently been proposed for scientific literature.

In the biomedical domain, datasets such as GENIA [3] and the ChEMU reaction corpus [4] extend event extraction to scientific and patent text. In materials science, existing studies mainly focus on synthesis procedure extraction and experimental workflow representation [5–8]. These studies commonly model events through local trigger expressions and associated argument spans, such as operations, materials, apparatuses, quantities, or reaction conditions.

Polymer biodegradation event extraction differs from these settings in both its scientific target and its required representation. Rather than describing a local operation such as adding, heating, or mixing as event triggers, a biodegradation event corresponds to an experimentally supported scientific finding: whether a polymer is degraded by a microorganism or enzyme under particular material and environmental conditions. Such findings may be supported by evidence distributed across materials, methods, and results sections, and therefore cannot always be anchored to a single local trigger expression.

In addition, the information needed to represent a biodegradation finding is not limited to entity mentions or short argument spans. Experimental conditions and observed outcomes may be expressed as longer textual descriptions, such as compost composition, pretreatment status, incubation settings, surface erosion, weight loss, or other degradation evidence. Therefore, a domain-adapted template is needed to represent not only event participants, but also experimentally important sample, condition, and result information.

A further distinction lies in the scientific interpretation of participant specificity and generalization. For polymers, experimentally specific material variants such as low-molecular-weight polyethylene '*LMWPE*' may be generalized to polyethylene '*PE*' according to constitutional repeating unit (CRU)-based normalization principles used in PoLyInfo [16]. This means that differences arising from molecular weight, preparation method, or higher-order material structure are abstracted away, while polymers are grouped by their molecular repeating-unit composition. For microorganisms, mentions such as '*Pseudomonas sp. E4*' and '*Pseudomonas*' reflect biologically meaningful taxonomic relations, such as strain-, species-, and genus-level specificity. Because closely related microorganisms may share genetic or metabolic properties relevant to biodegradation, such generalization is scientifically meaningful, but it must remain constrained by the evidence provided in the text.

These properties make polymer biodegradation event extraction substantially different from conventional trigger-centered event extraction and existing scientific workflow extraction tasks. Reliable representation and evaluation of such findings therefore require document-level event modeling together with condition-sensitive and specificity-aware scientific interpretation.

2.2. The PolyBD dataset

The PolyBD dataset proposed by Liu et al. (2025) [17] consists of 100 scientific papers on polymer biodegradation. All papers are manually curated into sentence-level annotations. Four domain-specific entity types are annotated: *Polymer*, *Bacterium*, *Fungus*, and *Enzyme*. For each entity mention, both the exact text span and character-level offsets are provided.

To preserve scientifically meaningful specificity information, PolyBD annotates mentions appearing at different specificity levels separately whenever such information is explicitly available in the text. For example, '*Pseudomonas sp. E4*' (strain), '*Pseudomonas sp.*' (species), and '*Pseudomonas*' (genus) are annotated as separate microorganism mentions. Similarly, specific polymer variants such as low-molecular-weight polyethylene '*LMWPE*' and broader polymer classes such as polyethylene '*PE*' are annotated as separate polymer mentions when both are explicitly supported by the text. Enzyme mentions are also annotated separately when nested or specificity-related expressions are available.

In total, the dataset contains 14,905 sentences and 18,947 annotated mentions.

3. The PolyBD-EE dataset

We construct PolyBD-EE, a document-level polymer biodegradation event extraction dataset built upon the PolyBD corpus. PolyBD-EE extends PolyBD with entity linking annotations, structured event annotations, and text-grounded event participant expansion. Dataset statistics are summarized in Tables 1 and 2.

Table 1. Statistics of the PolyBD-EE dataset. Examples of 'atomic event' and 'expanded entity/mention' are provided in Figure 1. 'Participant tuple' is constructed from combinations of atomic and expanded participants; refer to Section 4.3 for details and examples.

	Sentence	Word	Mention	Atomic Event	Expanded Entity/Mention	Participant Tuple
Average per document	186	26,140	244	20	30	71
Total (15 documents)	2784	392,098	3,653	304	454	1,070

Table 2. Statistics of participant arguments in PolyBD-EE dataset.

Argument	Atomic Event	Expanded Entity/Mention
Polymer	304	84
Microorganism	246	290
Enzyme	120	80
Total	670	454

3.1. Data preparation

Polymer biodegradation findings are often supported by experimental evidence distributed across multiple sections of a scientific article rather than localized within a single sentence or paragraph. Therefore, reliable event annotation requires access to full-text articles. We thus first verified paper accessibility through our institution before dataset construction. Among the 100 papers in PolyBD, 85 were accessible, including 33 open-access papers. We randomly selected 15 papers from these accessible papers for event annotation. We intentionally do not restrict the dataset to open-access papers (7 of 15) only. The dataset also includes 8 non-open-access papers, allowing the benchmark to cover biodegradation studies beyond literature that can be freely accessed online.

3.2. Structured representation of biodegradation findings

To represent experimentally grounded biodegradation findings in scientific literature, we construct a domain-adapted event template for polymer biodegradation, as shown in Table 3.

As discussed in Section 2.1, polymer biodegradation findings are often supported by experimentally distributed evidence extending beyond isolated triggers or local span-level representations. Therefore, PolyBD-EE does not annotate event triggers or define sentence-level boundaries for events.

Each event is organized around a biodegradation finding and must contain a *Polymer* participant together with at least one biodegradation-related participant from *Microorganism* or *Enzyme*. We additionally allow polymer-only events when biodegradation is explicitly reported but no responsible microorganism or enzyme is identified.

Besides participant arguments, we annotate contextual information including *Sample*, *Condition*, and *Result*. These arguments preserve material states, experimental settings, and degradation evidence associated with the biodegradation findings.

Each event is further assigned a binary *Polarity* attribute indicating whether biodegradation is reported to occur (Positive) or not occur (Negative). The inclusion of negative biodegradation findings is particularly important for materials informatics applications, since predictive models for biodegradability require both successful and unsuccessful biodegradation cases for training and evaluation. By explicitly representing experimentally supported non-degradation findings, PolyBD-EE provides information often omitted from biodegradation resources but essential for future biodegradability prediction studies.

Table 3. The biodegradation event template in the PolyBD-EE dataset.

Category	Role	Definition and Scope
Argument	*Polymer*	The target polymeric material undergoing biodegradation.
	Sample	Sample information of the polymer (e.g. molecular weight, oxidation state).
	Microorganism	Bacteria or fungi involved in degradation.
	Enzyme	Enzymes associated with polymer degradation.
	Condition	Experimental settings (e.g. temperature, pH, medium, incubation duration).
	Result	Experimental evidence supporting biodegradation (e.g. weight loss, gas evolution).
Attribute	*Polarity*	Binary label: **Positive** (degraded) or **Negative** (not degraded).

3.3. Data annotation

The annotation process consists of entity linking and event extraction. Basic annotations are manually curated by a single annotator with basic domain knowledge of polymer biodegradation and reviewed three times after the initial annotation pass. After further quality review on a subset of documents, some modifications were made to refine the final annotations (see Section 3.4).

3.3.1. Entity linking

Following the ACE2005 annotation strategy, we group mentions referring to the same real-world entity at the document level. Polymer entities are linked to PoLyInfo[1] [16], while bacterium and fungus entities are mapped to the NCBI Taxonomy Database[2] whenever possible. Entities not covered by these resources are assigned dataset-specific identifiers.

3.3.2. Event extraction

We annotate two levels of biodegradation events: *Atomic Events* and *Expanded Events*. Atomic events preserve experimentally specific biodegradation findings, while expanded events support higher-level aggregation and retrieval across related participants.

3.3.2.1. Atomic events. Atomic events represent the most specific biodegradation claims supported in the paper, including both experimentally validated findings and biodegradation relations cited from prior studies. Each atomic event is constructed around a participant set (polymer–microorganism–enzyme, polymer–microorganism, polymer–enzyme, or polymer only).

For example, if a paper reports that each of '*Pseudomonas sp. E4*' and '*recombinant E. coli 21*' can degrade '*LMWPE*' using '*alkane hydroxylase*', we annotate two separate atomic events corresponding to the two participant sets.

Condition and *Result* arguments are annotated using concise excerpts directly in the paper. When the same participant set is evaluated under different experimental conditions, the corresponding biodegradation findings are annotated as separate atomic events.

3.3.2.2. Expanded events. Expanded events extend atomic events using broader participant concepts explicitly supported by the textual evidence, without relying on external ontologies or implicit background knowledge. As discussed in Section 2.1, such expansion reflects scientifically meaningful participant generalization under constrained interpretation settings.

For example, for polymer participants, expansion may generalize experimentally specific material variants to broader polymer classes (e.g. '*LMWPE*' → '*PE*'); for microorganism participants, expansion may generalize strain- or species-level mentions to broader taxonomic concepts (e.g. '*Pseudomonas sp. E4*' → '*Pseudomonas sp.*').

However, expansion is only allowed when the generalized relation remains compatible with the original biodegradation claim. For example, the event based on ('*LMWPE*', '*recombinant E. coli 21*') can be expanded from '*LMWPE*' to '*PE*', but not from '*recombinant E. coli 21*' to '*E. coli*' because the text does not establish that the biodegradation capability generalizes from the recombinant strain to the broader organism concept.

3.4. Annotation quality review

To provide an additional quality check, a second researcher with expertise in polymer materials reviewed the annotations of five randomly selected articles against the original article text. The review covered all 62 initially annotated events in these articles. Of these, 56 events (90.3%) required no revision, while 6 (9.7%) were revised. Three revisions added a more specific polymer entity ('*polyester-based polyurethane*'), two added missing expanded microorganism annotations, and one corrected the granularity assignment between a microorganism and its expanded form. The review additionally identified two biodegradation events that had been missed during the original annotation. After verification against the source articles by the original annotator, all confirmed corrections were incorporated into the final gold annotations used for evaluation.

4. Contradiction-aware evaluation framework

Conventional event extraction evaluation typically assumes that predicted and gold events can be directly compared through exact or partial matching over triggers and argument spans. Under this assumption, event identity is largely determined by textual overlap or semantic similarity between localized event representations.

However, this assumption does not hold for document-level biodegradation findings. The same biodegradation finding may be described using different excerpts distributed across multiple parts of the document, causing events describing the same finding to share little lexical overlap.

A natural alternative is to compare events based on semantic similarity. However, similarity-based matching is also insufficient for biodegradation findings. Even when two events involve the same participants and highly similar experimental descriptions, small differences in degradation conditions or outcomes may indicate distinct findings. In such cases, high semantic similarity may incorrectly align incompatible events as the same biodegradation finding.

Therefore, event identity in biodegradation extraction cannot be reliably determined through either textual matching or similarity-based comparison alone. Instead, evaluation requires determining whether two structured events are mutually compatible as descriptions of the same biodegradation finding.

To investigate how such judgments are made in practice, we first conducted a human ranking study in which domain experts compared candidate event pairs and identified the predictions most consistent with a gold biodegradation finding (Section 4.1). Based on the resulting observations, we propose a contradiction-aware evaluation framework consisting of two perspectives: event identity evaluation (Section 4.2) and event-conditioned expansion evaluation (Section 4.3).

4.1. Preliminary human ranking study

Before designing the automatic evaluation framework, we first investigated how humans judge equivalence between polymer biodegradation events.

We use a simple extraction prompt to generate candidate biodegradation events from scientific papers.[3] For each gold event, three candidate predictions are selected and presented together with the gold annotation to a domain expert familiar with polymer biodegradation research. The expert ranks the candidate predictions according to how well they match the gold event and provides brief explanations for the ranking decisions. The study includes 52 event groups collected from 8 papers.

Our analysis shows that human judgment is primarily driven by whether two events describe the same

experimentally compatible biodegradation finding, rather than by textual or semantic similarity alone. Human evaluators first determine whether the compared events refer to the same biodegradation finding based on participant identity and polarity consistency, and then evaluate whether the findings remain compatible under the described experimental conditions/results. Contradictory experimental settings (e.g. pretreated vs. untreated materials, aerobic vs. anaerobic conditions, or soil-based vs. liquid-based environments) strongly reduce perceived event equivalence even when the compared descriptions remain textually or semantically similar.

Human evaluators also consistently prefer predictions that preserve finer-grained participant specificity when the text supported such information. For example, strain-level microorganism predictions are generally preferred over species- or genus-level predictions when the gold event is annotated at the strain level.

This preference reflects the scientific value of preserving specific information. More specific participant descriptions often retain information that can later support broader scientific interpretation. For polymers, specific material variants may provide evidence relevant to broader polymer classes sharing the same constitutional repeating unit. For microorganisms, biodegradation capability observed for a specific strain may also provide useful evidence for taxonomically related organisms.

Importantly, such a broader interpretation is often recoverable directly from the textual evidence itself, without requiring external ontology inference. As a result, predictions that fail to preserve finer-grained participant specificity are generally considered less informative during human evaluation, even when broader compatible concepts are correctly identified.

These observations motivate the contradiction-aware design of the proposed evaluation framework. Additional analysis is provided in Appendix A.

4.2. Event identity evaluation

Motivated by observations from the human ranking study, we perform event identity evaluation in three stages: participant identity gating, contradiction-based compatibility filtering, and global event alignment via maximum bipartite matching.

The proposed framework is not intended to provide aperfect definition of scientific event equivalence. Rather, our goal is to establish an initial contradiction-aware evaluation framework that better reflects how humans assess identity consistency between document-level biodegradation findings.

4.2.1. Pairwise identity gating

For each gold–prediction event pair, we first determine whether the prediction can be considered a valid candidate based on participants and polarity consistency. Under the *strict* setting (Figure 2(a)), participant arguments must exactly match the atomic arguments of the gold event. Under the *lenient* setting (Figure 2(b)), predicted participant arguments are additionally allowed to match expanded arguments derived from the gold event. The polarity of two events should be the same in both settings.

4.2.2. Contradiction-based compatibility filtering

After participant and polarity gating, we further determine whether two events are experimentally compatible. Contradiction detection is performed over both categorical experimental conditions and quantitative evidence extracted from event arguments.

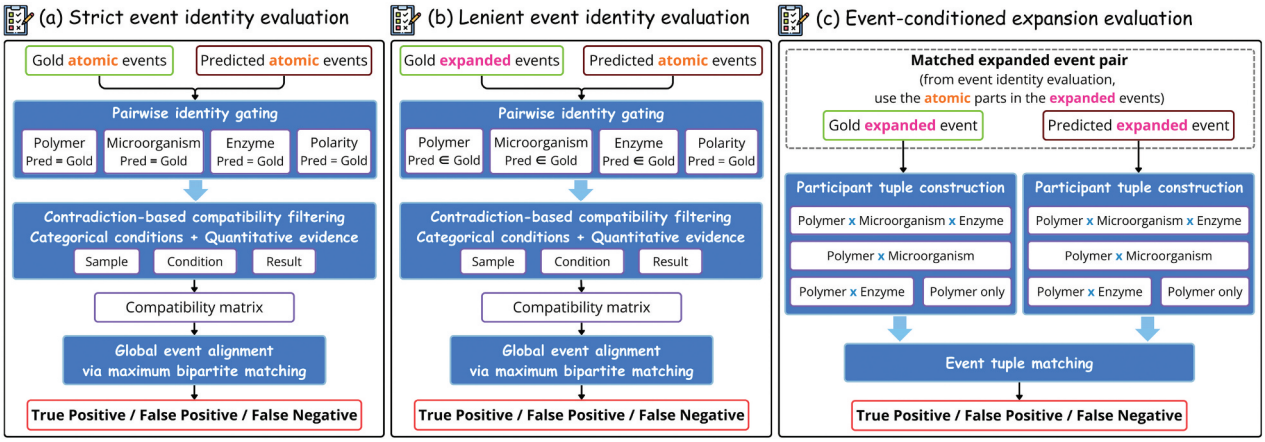


Figure 2. Illustration of the contradiction-aware evaluation framework for polymer biodegradation event extraction. (a) Strict event identity evaluation. Predicted atomic events are matched against gold atomic events using strict participant and polarity consistency, followed by contradiction-aware filtering and maximum bipartite matching. (b) Lenient event identity evaluation. Similar to the strict setting, except that participant matching additionally allows text-grounded expanded arguments derived from the gold event. (c) Event-conditioned expansion evaluation. Event identity matching is first performed using atomic arguments within expanded events. Expanded participant tuples are then generated from matched event pairs and compared using ID-level or mention-level matching.

For categorical experimental conditions, we define several event-defining dimensions, including pretreatment status, environmental setting, oxygen condition, culture setup, and organism setup. Event pairs expressing mutually incompatible conditions within the same dimension (e.g. aerobic vs. anaerobic, soil-based vs. liquid-based) are treated as contradictory and rejected during identity evaluation.

For quantitative evidence, we extract structured numerical information from event arguments, including temperature, incubation duration, pH, concentration, agitation speed, and other relevant measurements. Quantitative contradictions are evaluated only when both events explicitly provide values for the same category. Two quantitative conditions are considered compatible when at least one value pair falls within a predefined tolerance range.

Detailed contradiction rules, quantitative compatibility criteria, and extraction patterns are provided in Appendix B.1.

4.2.3. Global event alignment

After participant gating and contradiction-based compatibility filtering, we construct a bipartite graph between predicted and gold events.

Event pairs passing previous stages are treated as compatible alignments, while contradictory or incompatible pairs are excluded from further matching. We then perform maximum bipartite matching to obtain a globally consistent one-to-one alignment between predicted and gold events. For compatible event pairs, we assign a small secondary weight based on supporting quantitative evidence coverage. This weighting is used only for tie-breaking among already compatible candidate pairs.

A predicted event aligned with a gold event after global matching is treated as a true positive.

4.3. Event-conditioned expansion evaluation

As PolyBD-EE includes expanded-event annotations, we further provide a pipeline to evaluate the ability of models to generalize from atomic events to broader text-grounded concepts (see Figure 2(c)).

As described in Section 3.3.2, expanded events extend atomic events using broader participant concepts explicitly supported by the textual evidence, without relying on external ontologies or background knowledge. The expansion process applies to participant arguments, including polymers, microorganisms, and enzymes.

To evaluate expansion quality, we first perform event identity evaluation using the atomic arguments of the expanded events. For each successfully matched gold–prediction event pair, we then construct participant tuples from combinations of atomic and expanded participants.

For example, in *Event-1* of Figure 1, the expanded event contains 2 polymers, 3 microorganisms, and 2 enzymes, resulting in $2 \times 3 \times 2 = 12$ possible participant tuples, including the original atomic tuple (‘*LMWPE*’, ‘*Pseudomonas sp. E4*’, ‘*alkane hydroxylase*’). Predicted tuples are compared against gold tuples using participant matching: Entity IDs are used when available; enzyme participants are matched using mentions. A predicted tuple appearing in the gold tuple set is treated as a true positive.

We evaluate expansion using participant tuples rather than independent participant coverage because event expansion is fundamentally conditioned on the underlying atomic event. Expanding by one additional participant may generate multiple new event hypotheses due to combinatorial pairing between participants. Therefore, correct expansion requires not only predicting broader concepts, but also preserving valid event-level compatibility between expanded participants.

5. Zero-shot extraction with generative models

With the PolyBD-EE dataset and the proposed evaluation framework, we further investigate whether modern generative language models can perform polymer biodegradation event extraction under a zero-shot setting.

We focus on zero-shot extraction because PolyBD-EE is designed primarily as an evaluation benchmark rather than a training corpus. With only 15 documents, further splitting the dataset into training and test sets would leave very limited data for both model training and reliable evaluation. Few-shot prompting also presents a different challenge in this document-level setting, as each input consists of a full article together with its entity list, making the selection and inclusion of demonstration articles non-trivial. We therefore use zero-shot extraction as the primary setting for establishing an initial benchmark, while leaving supervised learning and systematic few-shot evaluation for future work.

5.1. Multi-turn zero-shot event extraction

Inspired by the human annotation process described in Section 3.3.2, we design a multi-turn conversational extraction framework for scientific event extraction. As illustrated in Figure 3, the framework consists of five sequential dialogue stages, where each stage retains the complete dialogue history from previous turns.

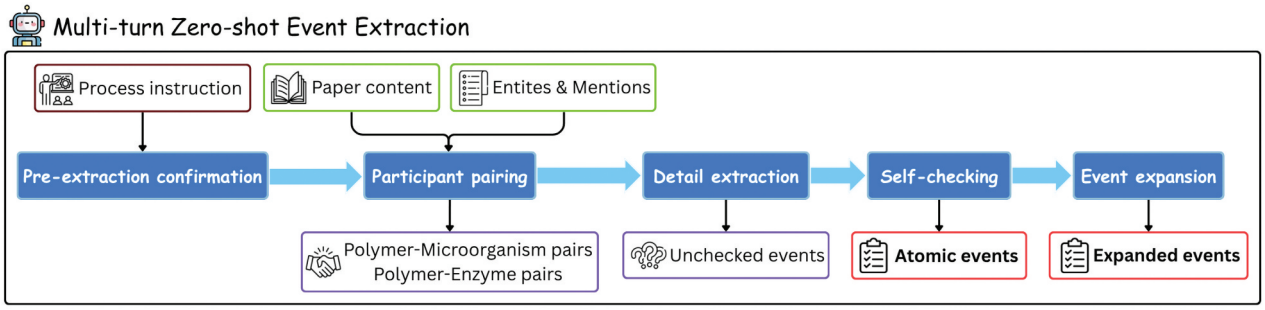


Figure 3. An illustration of the multi-turn zero-shot extraction method applying a generative language model on the polymer biodegradation event.

5.1.1. Pre-extraction confirmation

In the first stage, we provide the model with the task definition together with the objectives of the subsequent extraction stages, and ask the model to confirm its understanding of the task before extraction begins. We introduce this stage because polymer biodegradation event extraction involves complex document-level reasoning, including distributed evidence, multiple argument types, and excerpt-level evidence extraction. This stage is intended to align the model with the overall extraction objective before structured extraction starts.

5.1.2. Participant pairing

In the second stage, we provide the document text together with the pre-annotated entity and mention lists, and ask the model to identify biodegradation-related polymer–microorganism and polymer–enzyme pairs. Rather than directly extracting full event structures, we first decompose the task into binary participant relations. This decomposition separates participant association from later event-level construction and follows the staged annotation strategy used during dataset curation.

5.1.3. Detail extraction

In the third stage, the model is asked to construct event-level annotations based on the extracted participant pairs. Definitions of all event arguments are provided in the prompt, including special cases such as polymer-only events. The model then extracts detailed event information, including conditions, results, sample information, and polarity.

5.1.4. Self-checking

The model performs event-level deduplication and self-checking over the extracted atomic events to improve consistency and reduce redundant predictions.

5.1.5. Event expansion

In the final stage, the model expands the verified atomic events into expanded events by generating broader participant concepts supported by the textual surface form.

Overall, the multi-turn framework progressively decomposes scientific event extraction into increasingly complex subtasks while preserving complete dialogue history throughout the process. Earlier-stage predictions serve as structured intermediate references for later stages, which may help reduce cascading errors during event construction. All prompts used in the framework are provided in Appendix C.1.

5.2. Experimental setups

We evaluate six large language models on polymer biodegradation event extraction under a zero-shot setting. Both open-source and closed-source models are included.

For open-source models, we evaluate three models from the Qwen3.5 series (9B, 27B, and 122B) together with GPT-OSS-120B [18,19].[4] For closed-source models, we include GPT-5.4-mini[5] and GPT-5.56 as reference systems.

Each model is evaluated under two inference settings: *Instruct* and *Think*. In the *Instruct* setting, models directly generate the final JSON-formatted extraction results. In the *Think* setting, models are additionally allowed to generate intermediate reasoning content before producing the final JSON output.

All locally deployable models are evaluated on a single 80GB NVIDIA H100 GPU. Qwen-3.5-9B and Qwen-3.5-27B are deployed in BF16 format using the Transformers library. Qwen-3.5-122B and GPT-OSS -120B cannot be deployed in BF16 within the available GPU memory and are therefore deployed with Ollama using 4-bit quantization (Q4_K_M). Because the larger models require quantization under this hardware constraint, comparisons across model sizes may partly reflect differences in numerical precision rather than model scale alone. GPT-5.4-mini and GPT-5.5 are accessed through the official API. Additional implementation details are provided in Appendix C.2.

5.3. Evaluation metrics

We evaluate model performance using the event identity evaluation and event-conditioned expansion evaluation frameworks introduced in Section 4.

For atomic event extraction, we evaluate the events produced after the fourth extraction stage of the multi-turn framework. Precision, recall, and micro F1 are computed from the true positives, false positives, and false negatives obtained under both the strict and lenient event identity evaluation settings.

For event expansion evaluation, we use the expanded events generated during the fifth extraction stage. Precision, recall, and micro F1 are computed based on matching expanded participant tuples between predicted and gold expanded events.

Since expansion evaluation is performed only on event pairs successfully aligned during atomic event identity evaluation, the resulting metrics reflect conditional expansion performance rather than end-to-end extraction performance. To additionally measure overall end-to-end expansion recall, we report Overall Recall (O-Rec) and Overall F1 (O-F1). Unlike standard expansion recall and F1, which are computed only over tuples from successfully aligned gold events, O-Rec and O-F1 are computed over tuples derived from all gold events in the dataset, including those belonging to unmatched atomic events.

5.4. Main results

We evaluate model performance on two extraction objectives: atomic biodegradation event extraction and event-conditioned participant expansion. Specifically, we first evaluate whether models can correctly extract document-level atomic biodegradation events under the proposed contradiction-aware identity evaluation framework (RQ1). We then evaluate whether models can further expand these extracted events into broader text-grounded participant concepts (RQ2).

5.4.1. RQ1: can models extract polymer biodegradation events?

Table 4 reports the event identity evaluation results under both strict and lenient settings.

Overall, the best performance is achieved by GPT-5.5 under the *Think* setting, reaching 0.492 micro F1 under strict evaluation and 0.557 under lenient evaluation. However, the bootstrap confidence intervals are relatively wide and overlap across several models (e.g. 0.348–0.637 for GPT-5.5-*Think* under strict evaluation), reflecting the uncertainty associated with the limited benchmark size. Model rankings based on point estimates should therefore be interpreted cautiously.

Even though the proposed evaluation framework does not guarantee complete equivalence between matched predicted and gold events across the *Sample*, *Condition*, and *Result* arguments, substantial room for improvement remains before polymer biodegradation event extraction can be considered fully reliable.

We observe a consistent performance gap between strict and lenient evaluation across all models. This suggests that models are often able to identify the existence of a biodegradation event while failing to recover the finest-grained participant specificity described in the original paper. The absolute micro-F1 gap ranges from 2.7% for GPT-5.4-mini-*Instruct* to 11.9% for Qwen-3.5-27B-*Instruct*.

Reasoning-based generation generally improves extraction performance. All models achieve higher recall and micro F1 under the *Think* setting, with absolute strict-F1 improvements ranging from 3.1% to 18.4%. Paired bootstrap comparisons show statistically significant improvements for Qwen-3.5-9B (+18.4 points, $p < 0.001$), Qwen-3.5-122B (+16.2 points, $p = 0.036$), and GPT-5.5 (+9.0 points, $p = 0.015$), whereas the improvements for Qwen-3.5-27B, GPT-OSS-120B, and GPT-5.4-mini do not reach statistical significance. This observation suggests that intermediate reasoning can improve document-level biodegradation event extraction, but its benefit is model-dependent rather than uniform across systems.

Interestingly, GPT-OSS-120B exhibits the smallest gains (3.1% F1) under the *Think* setting. Although the

Table 4. Event identity evaluation results by several large language models on PolyBD-EE, with 95% bootstrap confidence intervals (document-level cluster bootstrap, $N = 2000$ resamples) reported for F1.

			Strict-Atomic				Lenient-Atomic			
Mode	Model	Avg. #Char of Thinking	Pre	Rec	F1	95% CI (F1)	Pre	Rec	F1	95% CI (F1)
Instruct	Qwen-3.5-9B	–	0.160	0.056	0.083	[0.042, 0.137]	0.274	0.095	0.142	[0.090, 0.212]
	Qwen-3.5-27B	–	0.292	0.194	0.233	[0.132, 0.354]	0.441	0.293	0.352	[0.221, 0.491]
	Qwen-3.5-122B	–	0.311	0.247	0.275	[0.143, 0.392]	0.386	0.306	0.341	[0.197, 0.478]
	GPT-OSS-120B	–	0.364	0.181	0.242	[0.133, 0.357]	0.523	0.260	0.347	[0.169, 0.506]
	GPT-5.4-mini	–	0.343	0.155	0.213	[0.111, 0.328]	0.387	0.174	0.240	[0.129, 0.374]
	GPT-5.5	–	0.405	0.401	0.403	[0.242, 0.546]	0.459	0.454	0.456	[0.296, 0.603]
Think	Qwen-3.5-9B	32,765	0.446	0.191	0.267	[0.171, 0.368]	0.562	0.240	0.336	[0.222, 0.461]
	Qwen-3.5-27B	12,676	**0.560**	0.322	0.409	[0.215, 0.592]	**0.623**	0.359	0.455	[0.264, 0.625]
	Qwen-3.5-122B	24,991	0.494	0.395	0.439	[0.265, 0.580]	0.568	0.454	0.505	[0.312, 0.662]
	GPT-OSS-120B	2,035	0.399	0.207	0.273	[0.125, 0.422]	0.475	0.247	0.325	[0.163, 0.476]
	GPT-5.4-mini	unknown	0.451	0.240	0.313	[0.188, 0.435]	0.549	0.293	0.382	[0.263, 0.487]
	GPT-5.5	unknown	0.465	**0.523**	**0.492**	[0.348, 0.637]	0.526	**0.592**	**0.557**	[0.408, 0.683]

same inference configuration is used, its generated reasoning traces are substantially shorter than those produced by Qwen-3.5-122B, potentially limiting the effectiveness of reasoning-guided extraction.

Substantial capability differences are also observed across model scales and families. Smaller models such as Qwen-3.5-9B remain far behind GPT-5.5, indicating that high-quality document-level scientific event extraction still requires considerable model capacity and computational resources. Moreover, even the strongest open-source model, Qwen-3.5-122B, remains more than 5% absolute F1 behind GPT-5.5 under both settings.

Under the *Think* setting, model precision becomes substantially more consistent across models. While precision varies considerably across models under the *Instruct* setting, all models exceed 39% precision under strict evaluation and 47% precision under lenient evaluation once intermediate reasoning is introduced. In contrast, the recall gaps observed under the *Instruct* setting largely persist under the *Think* setting. This suggests that reasoning traces may help stabilize event verification and reduce overly aggressive predictions across models, while the ability to recover additional valid events still strongly depends on the underlying model capability.

5.4.2. RQ2: can models correctly expand event participants?

Table 5 reports the event-conditioned expansion evaluation results.

GPT-5.5 under the *Think* setting achieves the highest overall F1, reaching 0.577 and 0.632 under strict and lenient event matching, respectively. However, its overall recall remains only 51.3% and 58.6%, indicating that even one of the strongest evaluated models still fails to recover a substantial proportion of the gold participant tuples in the dataset.

Interestingly, smaller models that perform poorly in event identity evaluation may still achieve relatively high precision on matched-event expansion evaluation. For example, Qwen-3.5-9B-*Think* and Qwen-3.5-27B-*Think* obtain higher strict-matched expansion precision than Qwen-3.5-122B-*Think* and GPT-5.5-*Think*. This phenomenon reflects a selection bias in matched-event evaluation: weaker models recover fewer atomic events, but those successfully recovered events are often relatively easier cases whose participant expansion is also simpler. This further demonstrates the importance of reporting expansion metrics computed over tuples from all gold events rather than only matched events.

Comparing model rankings across tasks also reveals that event identification and event expansion do not rely entirely on the same underlying capabilities. For example, under the *Think* setting, Qwen-3.5-27B obtains a higher point estimate than GPT-5.4-mini in event identity evaluation, whereas GPT-5.4-mini obtains higher expansion recall and overall F1. These pairwise differences do not reach statistical significance, but the reversal in point estimates illustrates that performance on event identification does not directly determine performance on participant expansion.

Finally, except for three smaller models (Qwen-3.5-9B/27B and GPT-5.4-mini), other models achieve over 92% recall on matched-event expansion evaluation under the strict-*Think* setting, while precision remains between 52.8% and 65.8%. This suggests that stronger models become increasingly capable of identifying broader concepts, but still tend to overgeneralize them into valid event participants. Therefore, event-conditioned expansion should not be treated as a simple lexical or nested-entity expansion problem. Instead, correct expansion requires learning an appropriate decision boundary for event-level compatibility between broader concepts and the underlying atomic event.

The paired-bootstrap analysis further highlights the distinction between conditional expansion quality and end-to-end expansion performance. For example,

Table 5. Event-conditioned expansion evaluation results by several large language models on PolyBD-EE. Precision, recall, and micro F1 are computed against gold tuples constructed from matched gold events. 'O-Rec' and 'O-F1' refer to the recall and F1 of gold tuples constructed from all gold events.

		Expansion (strict-atomic)					Expansion (lenient-atomic)				
Mode	Model	Pre	Rec	F1	O-Rec	O-F1	Pre	Rec	F1	O-Rec	O-F1
Instruct	Qwen-3.5-9B	0.406	0.206	0.274	0.012	0.024	0.433	0.259	0.324	0.027	0.051
	Qwen-3.5-27B	0.508	0.637	0.565	0.121	0.196	0.508	0.598	0.549	0.186	0.272
	Qwen-3.5-122B	0.496	0.833	0.622	0.167	0.250	0.487	0.756	0.592	0.221	0.304
	GPT-OSS-120B	0.552	0.837	0.665	0.120	0.197	0.407	0.754	0.529	0.151	0.221
	GPT-5.4-mini	0.733	0.737	0.735	0.131	0.222	0.651	0.715	0.682	0.143	0.234
	GPT-5.5	0.658	0.943	0.775	0.429	0.519	0.666	0.876	0.757	0.462	0.545
Think	Qwen-3.5-9B	0.749	0.498	0.598	0.120	0.206	0.518	0.452	0.483	0.136	0.215
	Qwen-3.5-27B	0.727	0.622	0.671	0.157	0.258	0.711	0.606	0.654	0.165	0.268
	Qwen-3.5-122B	0.587	0.926	0.718	0.385	0.465	0.598	0.882	0.713	0.427	0.498
	GPT-OSS-120B	0.528	0.931	0.674	0.114	0.188	0.516	0.779	0.621	0.138	0.218
	GPT-5.4-mini	0.658	0.847	0.741	0.218	0.327	0.632	0.785	0.700	0.255	0.364
	GPT-5.5	0.658	0.947	0.777	0.513	0.577	0.685	0.900	0.778	0.586	0.632

under the *Think* setting, GPT-5.5 and Qwen-3.5-122B do not differ significantly in matched-event expansion F1 under either strict or lenient evaluation, whereas GPT-5.5 achieves significantly higher O-F1 when all gold events are considered. This suggests that the overall advantage of GPT-5.5 arises not only from participant expansion itself, but also from stronger event-level coverage.

6. Discussion

Although the proposed event representation, evaluation framework, and extraction framework can be further extended in many directions, we focus our discussion on two empirical questions in this work.

First, we investigate whether the gold–prediction event pairs aligned under the proposed contradiction-aware evaluation framework indeed correspond to the same biodegradation finding. Since the framework is designed to approximate human event identity judgment, it is important to examine whether the resulting alignments remain scientifically consistent at the finding level.

Second, we examine the effectiveness of the self-checking stage in the multi-turn extraction framework. Recent extraction approaches increasingly rely on model self-reflection and self-correction mechanisms, yet their effectiveness for document-level scientific event extraction remains unclear. We therefore analyze whether the self-checking stage can reliably improve extraction quality under the proposed framework.

Finally, based on the limitations identified through the proposed evaluation framework and manual analysis, we discuss possible directions for future improvement.

6.1. Validation of evaluation framework

To further validate the proposed contradiction-aware evaluation framework, we manually examined both aligned and rejected predictions from the best-performing GPT-5.5-*Think* model.

Our analysis focuses on three questions: whether aligned gold–prediction event pairs indeed correspond to the same biodegradation finding, whether valid predictions are incorrectly rejected by the framework, and what insights the manually examined predictions provide regarding current challenges in document-level scientific event extraction.

6.1.1. Validation of accepted alignments

Under strict and lenient event identity evaluation, 154/159 (96.9%) and 174/180 (96.6%) aligned gold–prediction event pairs, respectively, were confirmed to describe the same biodegradation finding. These results indicate high precision of the framework's accepted alignments for the examined GPT-5.5-*Think* predictions.

Beyond validating event alignment, manual inspection revealed a recurring difference between gold annotations and model predictions, even when both described the same biodegradation finding. Gold annotations reorganize textual evidence into concise experimental components, whereas model predictions often preserve sentence-level evidence from the original document.

For example, one gold event decomposed the experimental setup into concise components including 'preoxidized polyethylene', 'cultivating at 50 °C for 2 weeks', and 'on the agar plate containing preoxidized polyethylene as a sole carbon source'. The corresponding prediction instead retained several sentences containing additional contextual descriptions and citations unrelated to the core biodegradation finding.

Preserving sentence-level evidence created two problems. First, information relevant to different event components was often not cleanly separated. Because the retained sentences simultaneously described experimental conditions and degradation outcomes, the same sentence fragments were frequently copied into both the *Condition* and *Result* fields rather than being decomposed into separate event components. Second, information belonging to different atomic events was often not cleanly separated. Even when the correct polymer and microorganism were identified, predictions frequently retained sentences describing multiple polymers or experimental instances. For example, a prediction centered on 'PE-1' still preserved result descriptions jointly discussing 'PE-1', 'PE-2', 'PE-3', and 'PE-4', whereas the corresponding gold event retained only the evidence directly associated with 'PE-1'.

6.1.2. Analysis of false rejections

We further manually examined false-positive predictions to determine whether valid event predictions were incorrectly rejected by the evaluation framework.

Under strict evaluation, 175 of 183 false positives were rejected during participant identity gating, indicating that their participant configurations could not be aligned to any gold event. Of the remaining eight predictions, three were rejected because of incompatible experimental conditions, while five passed compatibility filtering but were not selected during one-to-one bipartite matching. Manual inspection identified only one of the 183 false-positive predictions as a genuine false rejection, corresponding to a false-rejection rate of 0.55%. Among the other four predictions not selected during bipartite matching, one was correctly rejected, while three reflected event-granularity differences in which a single gold event

Table 6. Effect of the self-checking stage on strict-atomic event identity F1, evaluated via a paired bootstrap test (document-level cluster bootstrap, $N = 2000$ resamples, shared resample indices between the pre- and post-self-check evaluations). ΔF1 = F1(after)−F1(before).

Mode	Model	F1 (before)	F1 (after)	ΔF1	95% CI (ΔF1)	*p*-value
Instruct	Qwen-3.5-9B	0.080	0.083	0.005	[−0.046, +0.066]	0.915
	Qwen-3.5-27B	0.275	0.233	−0.041	[−0.097, −0.003]	**0.020**
	Qwen-3.5-122B	0.276	0.275	−0.001	[−0.004, +0.002]	0.628
	GPT-OSS-120B	0.271	0.242	−0.030	[−0.071, +0.005]	0.119
	GPT-5.4-mini	0.213	0.213	0.000	[−0.015, +0.018]	0.981
	GPT-5.5	0.411	0.403	−0.008	[−0.079, +0.049]	0.849
Think	Qwen-3.5-9B	0.253	0.267	0.015	[−0.062, +0.082]	0.656
	Qwen-3.5-27B	0.383	0.409	0.027	[−0.034, +0.099]	0.439
	Qwen-3.5-122B	0.433	0.439	0.005	[−0.021, +0.026]	0.597
	GPT-OSS-120B	0.269	0.273	0.003	[−0.015, +0.023]	0.830
	GPT-5.4-mini	0.376	0.313	−0.060	[−0.180, +0.014]	0.334
	GPT-5.5	0.482	0.492	0.011	[−0.010, +0.034]	0.414

was represented as two predicted events containing different parts of the same finding.

Under lenient evaluation, 149 of 162 false-positive predictions failed participant identity gating. Of the remaining 13 predictions, four were rejected because of incompatible experimental conditions, and nine were not selected during bipartite matching. Manual inspection identified three genuine false rejections, corresponding to a false-rejection rate of 1.85% (3/162). Among the other six predictions not selected during bipartite matching, three were correctly rejected, and three reflected the same event-granularity difference described above.

These results indicate that most false positives arise from participant configurations that cannot align to any gold event, while the evaluation framework rarely rejects incorrect predictions.

6.1.3. Diagnostic analysis of model errors

Taken together, the manual analysis suggests that the main limitations of current model predictions occur at two different stages. First, incorrect participant configurations are a major source of false positive predictions. Second, even among predictions that are successfully aligned to gold events and therefore describe the same biodegradation finding, information belonging to different event arguments or different atomic events may remain mixed within the predicted event.

These findings indicate that current frontier models have not yet reached reliable document-level biodegradation event extraction. Recent evaluations of advanced AI systems have similarly shown that strong performance on many established benchmarks can coexist with substantial limitations on complex scientific tasks [20,21]. Our results suggest that document-level biodegradation event extraction remains a challenging problem with significant room for improvement. In this context, task-specific datasets and evaluation frameworks remain important for measuring progress, diagnosing failure modes, and guiding future scientific event extraction systems.

6.2. Effectiveness of self-checking

In the proposed multi-turn extraction framework (Section 5.1), we introduce a self-check stage to improve event consistency and reduce redundant predictions. To evaluate its effectiveness, we compare strict event identity evaluation results before and after the self-check stage using paired document-level bootstrap tests. Specifically, we use the outputs generated at the third stage as predictions, without self-checking, and compare them with the fourth-stage outputs.

As shown in Table 6, the self-check stage does not consistently improve atomic event extraction performance. Under the *Instruct* setting, self-checking generally leads to performance degradation or negligible changes, with Qwen-3.5-27B showing a statistically significant decrease of 4.1 F1 points ($p = 0.02$). Under the *Think* setting, positive effects are observed for most models, although the absolute micro-F1 gains remain relatively small, ranging from 0.3 to 2.7 F1 points; none of these improvements are statistically significant.

In particular, GPT-5.4-mini-*Think* exhibits an absolute micro-F1 decrease of 6.0 points after self-checking, although this difference is not statistically significant ($p = 0.334$). Since intermediate reasoning content is not exposed for this API-based model, the underlying cause cannot be directly inspected. One possible explanation is that the self-check stage may implicitly encourage more conservative prediction behavior, causing some models to remove previously extracted events more aggressively during revision.

The self-check stage involves event deduplication, consistency verification against source documents, and argument-level validation. In principle, such operations are expected to primarily improve precision by removing inconsistent or redundant event predictions. However, the largest observed gain is only 2.7% micro-F1 under Qwen-3.5-27B-*Think*, while the two strongest models, Qwen-3.5-122B-*Think* and GPT-5.5-*Think*, yield only 0.5% and 1.1% micro-F1 improvements, respectively.

These observations suggest that effective self-checking remains challenging for scientific event extraction, and more specialized verification mechanisms may be required. More importantly, these results demonstrate the value of task-specific datasets and evaluation frameworks for systematically assessing improvement strategies. Overall, the proposed framework enables quantitative assessment of techniques such as self-checking and provides a reliable basis for evaluating future advances in scientific event extraction systems.

6.3. Future directions

Although the self-checking strategy did not consistently improve extraction performance, the error patterns identified in Section 6.1 suggest two possible directions for improving event extraction.

First, one possible direction is to improve participant-level event hypothesis generation before detailed event construction. Since candidate polymers, microorganisms, and enzymes are already identified before event construction, a practical approach may be to enumerate candidate participant sets and prune them using smaller supervised filtering models before detailed event generation.

Second, models may benefit from mechanisms for generating concise event-specific summaries rather than directly preserving large sentence fragments from the source document. Because the proposed contradiction-aware evaluation framework evaluates event compatibility rather than exact textual overlap, it does not require event arguments to be copied directly from the source document and, therefore, naturally supports summarized or rewritten evidence representations.

A further direction concerns annotation validation. Although the annotation quality review in Section 3.4 provides an independent expert assessment of a subset of PolyBD-EE, it is not equivalent to independent annotation by a second annotator, and inter-annotator agreement could therefore not be assessed in this study. Future work should involve independent annotation of additional documents by multiple annotators to further evaluate the reproducibility of the proposed annotation scheme.

More broadly, the analyses in Sections 6.1 and 6.2 demonstrate that the proposed evaluation framework can provide diagnostic information beyond aggregate performance scores, helping identify specific bottlenecks that limit current scientific event extraction systems and guiding future system development.

7. Conclusion

In this work, we introduced PolyBD-EE, a document-level dataset for polymer biodegradation event extraction featuring structured event annotations and text-grounded participant expansion. We further proposed a contradiction-aware evaluation framework for experimentally grounded scientific events, enabling evaluation of both event identity and event-conditioned participant expansion.

Using PolyBD-EE, we conducted systematic zero-shot evaluations of modern generative large language models. Experimental results showed that current models can capture the underlying biodegradation findings while still struggling to recover the finest-grained participant specificity described in scientific literature. We further observed that models may over-generalize broader participant concepts during event expansion, and self-checking mechanisms provide only limited and inconsistent improvements. Through the proposed evaluation framework and manual analysis, we identified several bottlenecks that can guide future system development, including incorrect participant configurations, weak separation of event-specific information, and limitations of current self-checking strategies.

Overall, our findings indicate that document-level biodegradation event extraction remains a challenging problem with substantial room for improvement. We hope that PolyBD-EE and the proposed contradiction-aware evaluation framework will provide not only a benchmark for measuring future progress, but also a foundation for diagnosing failure modes and guiding the development of more reliable scientific event extraction systems in materials science and related domains.

Notes

1. https://polymer.nims.go.jp/.
2. https://www.ncbi.nlm.nih.gov/taxonomy.
3. GPT-4o-mini was used to generate predictions.
4. All four models were downloaded from HuggingFace in April 2026.
5. Model name: gpt-5.4-mini-2026–03-17.
6. Model name: gpt-5_5-2026–04-23.

Acknowledgements

The authors disclose that large language models and other AI-assisted tools were used only in a limited supporting role during manuscript preparation, including grammar and spell checking, language translation assistance, and support for code documentation. AI-based methods were also used where relevant in the research methodology and evaluation. All scientific decisions, analyses, and interpretations were made and verified by the authors.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was partially supported by the Data Creation and Utilization-Type Material Research and Development Project [Grant Number JPMXP1122714694].

ORCID

Shanshan Liu http://orcid.org/0009-0003-4370-4737
Yuji Matsumoto http://orcid.org/0000-0003-4946-9574
Masashi Ishii http://orcid.org/0000-0003-0357-2832

References

[1] Walker N, Trewartha A, Huo H, et al. The impact of domain-specific pre-training on named entity recognition tasks in materials science. Available at SSRN 3950755. 2021.
[2] Walker C, Strassel S, Medero J, et al. Ace 2005 multilingual training corpus. 2006:LDC2006T06. doi: 10.35111/mwxc-vh88
[3] Thompson P, Nawaz R, McNaught J, et al. Enriching a biomedical event corpus with meta-knowledge annotation. BMC Bioinf. 2011;12(1):393. doi: 10.1186/1471-2105-12-393
[4] He J, Nguyen DQ, Akhondi SA, et al. Chemu 2020: natural language processing methods are effective for information extraction from chemical patents. Front Res Metr Anal. 2021;6 - 2021. Available from: https://www.frontiersin.org/journals/research-metrics-and-analytics/articles/10.3389/frma.2021.654438
[5] Mysore S, Jensen Z, Kim E, et al. The materials science procedural text corpus: annotating materials synthesis procedures with shallow semantic structures. In: Friedrich A, Zeyrek D Hoek J, editors. Proceedings of the 13th linguistic annotation workshop. Florence (Italy): Association for Computational Linguistics; 2019 Aug. p. 56–64. Available from: https://aclanthology.org/W19-4007/
[6] Wang Z, Kononova O, Cruse K, et al. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. Sci Data. 2022;9(1):231. doi: 10.1038/s41597-022-01317-2
[7] Yang X, Zhuo Y, Zuo J, et al. PcMSP: a dataset for scientific action graphs extraction from polycrystalline materials synthesis procedure text. In: Goldberg Y, Kozareva Z Zhang Y, editors. Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi (United Arab Emirates): Association for Computational Linguistics; 2022 Dec. p. 6033–6046. Available from: https://aclanthology.org/2022.findings-emnlp.446/
[8] He S, Peng X, Cai Y, et al. ZSEE: a dataset based on zeolite synthesis event extraction for automated synthesis platform. In: Duh K, Gomez H Bethard S, editors. Findings of the Association for Computational Linguistics: NAACL 2024. Mexico City (Mexico): Association for Computational Linguistics; 2024 Jun. p. 1791–1808. Available from: https://aclanthology.org/2024.findings-naacl.116/
[9] Yoon MG, Jeon HJ, Kim MN. Biodegradation of polyethylene by a soil bacterium and alkb cloned recombinant cell. J Bioremed Biodegrad. 2012;3:1–8. Available from: https://api.semanticscholar.org/CorpusID:21693954
[10] Kumar R, Verma A, Shome A, et al. Impacts of plastic pollution on ecosystem services, sustainable development goals, and need to focus on circular economy and policy interventions. Sustainability. 2021;13(17):9963. Available from: https://www.mdpi.com/2071-1050/13/17/9963
[11] Mihai FC, Gündoğdu S, Markley LA, et al. Plastic pollution, waste management issues, and circular economy opportunities in rural communities. Sustainability. 2022;14(1):20. Available from: https://www.mdpi.com/2071-1050/14/1/20
[12] Patrício Silva AL. Future-proofing plastic waste management for a circular bioeconomy. Curr Opin Environ Sci Health. 2021;22:100263. Available from: https://www.sciencedirect.com/science/article/pii/S2468584421000350
[13] Ferreira-Filipe DA, Paço A, Duarte AC, et al. Are biobased plastics green alternatives?—A critical review. Int J Environ Res Public Health. 2021;18(15):7729. Available from: https://www.mdpi.com/1660-4601/18/15/7729
[14] SAPEA. Biodegradability of plastics in the open environment. Science Advice for Policy by European Academies; 2020. Available from: https://www.sapea.info/topics/biodegradability-of-plastics
[15] Afshar SV, Boldrin A, Astrup TF, et al. Degradation of biodegradable plastics in waste management systems and the open environment: a critical review. J Clean Product. 2024;434:140000. Available from: https://www.sciencedirect.com/science/article/pii/S0959652623041586
[16] Ishii M, Ito T, Sado H, et al. Nims polymer database polyinfo (i): an overarching view of half a million data points. Sci Technol Adv Mater. 2024;4(1):2354649. doi: 10.1080/27660400.2024.2354649
[17] Liu S, Ishii M, Matsumoto Y. Zero-shot entity recognition for polymer biodegradability information: gpt-4o on polybd. In: Proceedings of the thirty-first annual meeting of the association for natural language processing. Kyoto, Japan: Association for Natural Language Processing; 2025 Mar. p. 2377–2382. Available from: https://www.anlp.jp/proceedings/annual_meeting/2025/pdf_dir/P6-7.pdf
[18] Qwen Team. Qwen3.5: towards native multimodal agents. 2026. Available from: https://qwen.ai/blog?id=qwen3.5
[19] OpenAI. Gpt-oss-120b and gpt-oss-20b model card. 2025. Available from: https://arxiv.org/abs/2508.10925
[20] Phan L, Gatti A, Li N, et al. A benchmark of expert-level academic questions to assess ai capabilities. Nature. 2026;649(8099):1139–1146. Available from: https://www.nature.com/articles/s41586-025-09962-4
[21] Huang Z, Cao Y, Shargh AK, et al. Can coding agents reproduce findings in computational materials science? 2026. Available from: https://arxiv.org/abs/2605.00803

Appendix A. Observation of human ranking study

Our analysis shows that human judgment is primarily driven by event identity consistency rather than shallow semantic overlap. Human evaluators tend to determine whether two events describe the same biodegradation finding under compatible experimental settings, instead of simply comparing overlapping arguments or textual similarity.

A.1 Event identity consistency

The first and most important criterion is whether the prediction refers to the same biodegradation relation as the gold event. Human judges primarily focus on polymer identity, microorganism identity, and polarity consistency. Predictions inconsistent with these core identities are typically ranked lowest, even when other arguments partially overlap.

A.2 Experimental outcome consistency

When the polymer–microorganism–polarity identities are consistent, judges further examine whether the prediction describes the same type of experimental outcome. For example, weight loss, CO_2 evolution, surface erosion, and molecular-weight reduction are treated as different forms of biodegradation evidence. Predictions describing different result types are usually considered mismatched, even if the involved participants remain identical.

A.3 Condition consistency and experimental precision

Human judges additionally compare the experimental conditions associated with each event. Contradictory conditions, such as different pretreatment states, environments, incubation settings, or experimental branches, substantially reduce perceived event equivalence. In contrast, predictions with missing numeric details may still receive relatively high rankings as long as no explicit contradiction exists. When multiple predictions are conditionally compatible with the gold event, those containing more precise experimental descriptions are generally preferred.

A.4 Granularity preference

Human evaluators also show a clear preference for finer-grained entity specificity. For example, when the gold annotation of the Microorganism is defined at the strain level, predictions at the strain level are preferred over species-level predictions, while species-level predictions are preferred over genus-level predictions. This suggests that entity granularity plays an important role in human judgments of biodegradation event identity.

Appendix B. Details of evaluation framework

B.1 Contradictory-based filtering

The second stage of event identity evaluation focuses on contradiction-based compatibility filtering over experimental conditions and outcomes. We perform contradiction

Table B1. Experimental dimensions and contradiction states used in event identity evaluation. for each event-defining dimension, we detect whether the gold and predicted events express mutually exclusive experimental states. if both sides contain confidently identifiable but incompatible states within the same dimension, the event pair is rejected during identity evaluation.

Dimension	State	Representative Expressions
Pretreatment	Irradiated/Pretreated	Irradiated, UV, photo-oxidized, oxidized, pretreated, modified
	Untreated	Untreated, non-irradiated, not irradiated
Environment	Soil-based	Soil, soil burial, landfill
	Liquid-based	Liquid, medium, broth, flask, culture medium
Oxygen Condition	Aerobic	Aerobic, air, oxygen
	Anaerobic	Anaerobic, oxygen-free, nitrogen atmosphere, N2 atmosphere, inert atmosphere
Culture Setup	Shaking	Shaking, rpm, rotary shaker
	Static	Static culture, without shaking
Organism Setup	Consortium	Consortium, mixed culture, microbial community
	Isolate	Pure culture, isolate, strain

Table B2. Quantitative evidence extraction patterns and compatibility criteria used for contradiction-based event identity evaluation.

Category	Example Expressions	Compatibility Rule
Temperature	30 °C, 37°C	At least one value pair differs by ≤1 °C
Time (hours)	24 h, 48 hours	At least one value pair differs by ≤1 hour
Time (days)	30 days, ca. 30 days	At least one value pair differs by ≤1 day
Time (weeks)	2 weeks	At least one value pair differs by ≤0.5 week
Time (months)	3 months	At least one value pair differs by ≤0.5 month
Time (years)	1 year	At least one value pair differs by ≤0.1 year
pH	pH 7, pH 6.5–7.0	At least one value pair differs by ≤0.2
Concentration	1 g/L, 0.5 g L-1	At least one value pair differs by ≤0.1 g/L
Percentage	5%	At least one value pair differs by ≤1%
Agitation speed	150 rpm	At least one value pair differs by ≤20 rpm
Volume	50 mL	At least one value pair differs by ≤1 mL
ppm	100 ppm	Exact match required

detection over both categorical experimental states and quantitative experimental evidence extracted from event arguments.

For categorical states, we define several event-defining experimental dimensions, including pretreatment status, environmental setting, oxygen condition, culture setup, and organism setup. Each dimension contains mutually exclusive states (e.g. *aerobic* vs. *anaerobic*). If both the gold and predicted events explicitly express different states within the same dimension, the event pair is treated as contradictory and rejected during identity evaluation. The predefined state dimensions and corresponding contradiction rules are summarized in Table B1.

For quantitative evidence, we use regular expressions to extract structured numerical information from event arguments, including temperature, incubation duration, pH, concentration, agitation speed, percentage, volume, and ppm values. Contradictions are evaluated only when both the gold and predicted events explicitly provide values for the same quantitative category. Quantitative values are considered compatible when at least one gold–prediction value pair falls within a predefined tolerance range. The quantitative evidence categories and compatibility criteria are summarized in Table B2.

In addition, for event pairs that pass contradiction filtering, we compute a supporting evidence score reflecting the proportion of quantitative evidence categories in the gold event that are also compatibly supported by the prediction. This score is used only as a secondary weighting factor during global event alignment and does not override contradiction-based rejection.

B.2 Global event alignment

After gating and contradiction filtering, we construct a weighted compatibility matrix between predicted and gold events. Event pairs that pass both stages receive a score of $1 + 0.1\times$ support score, while incompatible pairs receive a score of 0. We then perform maximum bipartite matching over the compatibility matrix. A predicted event successfully matched to a gold event is treated as a true positive.

Appendix C. Experimental details

C.1 *Prompts for event extraction*

Prompts for each stage of the proposed multi-turn zero-shot extraction framework are shown in Figures C1–C5.

C.2 *Model setups*

For all evaluated models, the temperature is set to 0.7. For all locally implemented models, we set `top_p = 0.95, top_k = 20, repetition_penalty = 1.0`.

For Qwen-3.5-9B and Qwen-3.5-27B, `max_new_tokens` is set to 32,768 in order to avoid truncation of long intermediate reasoning content and structured extraction outputs. For all other models, the default output length settings provided by the corresponding inference frameworks or APIs are used.

The *Instruct* and *Think* inference settings are implemented differently depending on the deployment framework and model family. For Qwen-3.5-122B and GPT-OSS-120B deployed with Ollama, the `think = True` option is enabled in the *Think* setting. For Qwen-3.5-9B and Qwen-3.5-27B deployed using the Transformers library, the <think> token is appended to the end of the model input to trigger reasoning generation. For GPT-5.4-mini and GPT-5.5 accessed through the OpenAI API, the *Think* setting is implemented using `reasoning={'effort': 'medium'}`, and the *Instruction* setting is implemented using `reasoning={'effort': 'none'}`.

Prompt for the task setting and pre-extraction confirmation step:
You are an expert in polymer biodegradability. You are reading a paper and attempting to extract all biodegradation event information from the paper in a structured format. The document-level extraction should rely only on textual evidence, not on ontology expansion.
You divide the extraction process into four steps.
- **Step 1:** Pair given polymers with the microorganisms (including bacteria and fungi) mentioned in the paper. Based on the paper's content, identify the most specific (finest-grained) one-to-one (polymer, microorganism) pairs that are supported by at least one biodegradation event in the paper (including negative/absence results where the polymer was tested with the microorganism but not degraded).
For example, if the paper mentions that microorganisms A and B can each potentially give rise to polymer C, you need to construct two pairs of relationships: (C, A) and (C, B).
Additionally, if an enzyme is responding to a polymer's biodegradation, but no specific microorganism that produces the enzyme is mentioned, you should add the most specific (finest-grained) one-to-one relationships between the polymer and the enzyme into candidate pairs. Noticed that, literature examples cited in the paper (even not in the experiment section) also should be extracted, even if no experimental details (conditions/results) are provided in the current paper.
- **Step 2:** After identifying candidate pairs (one-to-one relationships between polymer and microorganism (or polymer and enzyme)), based on the paper's content and the provided enzyme mention list, identify the enzyme(s) that act in each polymer-microorganism event (you can pass the enzyme identification for polymer-enzyme candidate pairs), as well as the sample information of the polymer, the specific experimental conditions, the detailed results, and polarity for each event. The condition and result must be an excerpt from the paper (this can include multiple, concise, precise, accurate, appropriate, non-consecutive excerpts). And the polarity is a single-choice question with two options: Positive (degraded) or Negative (not degraded). If wording is indirect (e.g., “weight loss observed”, “surface erosion”), you should classify it as Positive, unless explicitly negated. If the paper includes degradations of a pair of polymer and microorganism (or polymer and enzyme) under different conditions (e.g., time, temperature, pH, etc.), you should treat the events under different conditions as multiple events. Additionally, for each polymer-enzyme candidate pair, you also need to identify the specific experimental conditions, sample information of the polymer, detailed results, and polarity for this pair to construct an event.
- **Step 3:** You should self-check all extracted events. This step includes removing duplicate events, confirming that the information in the event is consistent with the paper, and ensuring that the specific ID information of all entities (polymer and microorganism), the mentions (polymer, microorganism, enzyme), the excerpts (sample, condition, result), and the polarity in the event are correct. No need for identifying explicit trigger words for events.
- **Step 4:** You are expanding extracted biodegradation events using only text-grounded evidence (content of the paper). Given the specific events you extracted in the previous step, expand them only if the generalization is directly supported by how the terms are written in the text. Note that you do not use external knowledge or ontology. The expansion includes polymer, microorganism, and enzyme arguments. Given a polymer in the extracted event, if there is another polymer entity that can be generalized to (e.g., LDPE → PE), and the generalization can be directly observed from the text, you can expand an event to include the broader concept. Given a microorganism entity or enzyme mention (e.g., Marinobacter hydrocarbonoclasticus → Marinobacter; Lipase Asahi → Lipase), similar to the rule on the polymer entity expansion, as long as the connection is clear from the text surface, you can expand. The expansion is not limited to one-hop; you can keep expanding until there is no relation between the specific entity/mention and a broader concept can be observed. Since the list of polymer entities, the list of microorganism entities, and the list of enzyme mentions are provided at the beginning, the provided entities/mentions should be considered first during the expansion.
After these four steps, you have completed your analysis of a paper.

Before starting the extraction, you've already organized the required paper content and entity/mention information.
- The paper content is a list, with each item corresponding to a sentence in the paper.
- The polymer information is a dictionary of entities; the keys are specific IDs for the polymer, while the values are a list of all corresponding mentions of the polymer in the paper.
- The microorganism information is a dictionary of entities; the keys are specific IDs for the microorganism, while the values are a list of all corresponding mentions of the microorganism in the paper.
- The enzyme information is a list of multiple mentions, and you'll receive a list. Each item in the list represents all mentions of the enzyme class in the paper. You don't have the corresponding specific ID for each mention. Polymers mentioned with passive degradation, but no specific microorganism or enzyme is involved, are not targeted information.
An event must include a polymer entity, and one of a microorganism or an enzyme. No need for identifying explicit trigger words of events.
The output format of each step will be provided after the pre-extraction confirmation.
Before you start analyzing a paper, confirm what you need to do (i.e., what information you will obtain and what the extraction steps are), as well as what you are still unclear about. So, please start the pre-extraction confirmation.

Figure C1. The prompt for the first turn of our multi-turn zero-shot event extraction method.

Prompt for the pairing step:
Great. Now you receive the information for step 1.
Provided info:
Paper content: {paper_content}
Polymer entities: {polymer_content}
Microorganism entities: {micro_content}
Enzyme mentions: {enzyme_content}

Output format
You need to output candidate pairs by filling in the placeholders in [...] in a JSON format:

```
{
 "pair 1": {
  "Polymer": {"id": "...", "mention": "..."},
  "Microorganism": {"id": "...", "mention": "..."},
  "Enzyme": {"mention": "..."}
 },
 "pair 2": {
  "Polymer": {"id": "...", "mention": "..."},
  "Microorganism": {"id": "...", "mention": "..."},
  "Enzyme": {"mention": "..."}
 },
}
```

Notes
If there is a polymer-microorganism pair, you can set the value of "Enzyme" as {"mention": "None"}. Similarly, given a polymer-enzyme pair, you can set the value of "Microorganism" as {"id": "None", "mention": "None"}. The value of "mention" should be a list, even if there is only one mention proper.
If no measured biodegradation events concern the provided polymer entities, you can proceed with output and only output "The provided polymer entities are not involved in any event" in step 1.
In this turn, you only need to do step 1. Please start the Pairing (Step 1).

Answer

Figure C2. The prompt for the second turn of our multi-turn zero-shot event extraction method.

Prompt for the event-detail step:
Great. Now you have the extracted candidate pairs, and the list of enzyme mentions is provided in step 1. You can proceed to step 2.
Sometimes, polymer biodegradation occurs without specifying the specific microorganism involved (e.g., buried in soil; in vivo; etc.). If a polymer is biodegraded or not biodegraded under such conditions, the event involving the polymer and the corresponding conditions should also be extracted.

Output format
You need to output events by filling in the placeholders in [...] in a JSON format:

```
{
 "event 1": {
   "Polymer": {"id": "...", "mention": "..."},
   "Microorganism": {"id": "...", "mention": "..."},
   "Enzyme": {"mention": "..."},
   "Sample": {"excerpt": "..."},
   "Condition": {"excerpt": "..."},
   "Result": {"excerpt": "..."},
   "Polarity": ["Positive" / "Negative"]
 },
}
```

Notes
If there is event without enzyme involvement, you can set the value of "Enzyme" as {"mention": "None"}. Similarly, for a event without microorganism involvement, you can set the value of "Microorganism" to {"id": "None", "mention": "None"}. The value of "mention" and "excerpt" should be a list, even if there is only one mention proper. For "Polarity", the value should be a string either "Positive" or "Negative".
If no measured biodegradation events concern the provided polymer entities, you can proceed with output and only output "The provided polymer entities are not involved in any event" in step 2. In this turn, you only need to do step 2.
Please start the Event details (Step 2).

Answer

Figure C3. The prompt for the third turn of our multi-turn zero-shot event extraction method.

Prompt for the self-check step:
Great. Now you have the extracted event list. You can proceed to step 3.
In the previous step, you extracted the event information from the paper. Now, you need to further examine the extracted information and organize it. You need to output the event list with the format exactly as you used in step 2.

Notes
Don't forget the specific requirements we set for the self-check step; do your best to do so, and output only the event information in the given format.
If no measured biodegradation events concern the provided polymer entities, you can proceed with output and only output "The provided polymer entities are not involved in any event" in step 3. In this turn, you only need to do step 3.
Please start the Self-check (Step 3).
Answer

Figure C4. The prompt for the fourth turn of our multi-turn zero-shot event extraction method.

Prompt for the event-expansion step:
Great. Now you have the checked extracted event list. You can proceed to step 4.
After previous steps, you extracted the finest-grained event information from the paper. Now, you need to expand extracted events and organize them using only text-grounded evidence (content of the paper).
Output format
You need to output events by filling in the placeholders in [...] in a JSON format:

```
{
 "event 1": {
   "Polymer": {"id": "...", "mention": "..."},
   "Expanded_Polymer": [
     {"id": "...", "mention": "..."},
     {"id": "...", "mention": "..."}
   ],
   "Microorganism": {"id": "...", "mention": "..."},
   "Expanded_Microorganism": [
     {"id": "...", "mention": "..."},
     {"id": "...", "mention": "..."}
   ],
   "Enzyme": {"mention": "..."},
   "Expanded_Enzyme": {"mention": "..."},
   "Sample": {"excerpt": "..."},
   "Condition": {"excerpt": "..."},
   "Result": {"excerpt": "..."},
   "Polarity": ["Positive" / "Negative"]
 },
}
```

Notes
If there is an event without a border polymer entity or microorganism entity, you can set the value of "Expanded_Polymer" or "Expanded_Microorganism" as [{"id": "None", "mention": "None"}]. Similarly, for an event without a border enzyme mention, you can set the value of "Expanded_Enzyme" to {"mention": "None"}. The value of "mention" and "excerpt" should be a list, even if there is only one proper mention.
Don't forget the specific requirements we set for the expansion step; do your best to do so, and output only the event information in the given format.
If no measured biodegradation events concern the provided polymer entities, you can proceed with output and only output "The provided polymer entity is not involved in any event" in step 4. In this turn, you only need to do step 4.
Please start the Event Expansion (Step 4).

Answer

Figure C5. The prompt for the fifth (and last) turn of our multi-turn zero-shot event extraction method.