From Facts to Hypotheses: Joint Detection of Biomedical Relations and Epistemic Commitment Using LLMs

Aleksandra Gabryszak*, Phuc Tran Truong*, Arne Binder*
Nikola Milošević†, Felix-Sebastian Keese†, Astrid Rheinländer†, Philippe Thomas*

*German Research Center for Artificial Intelligence (DFKI), Salzufer 15/16, 10587 Berlin
†Bayer AG, Müllerstraße 178, 13353 Berlin
{aleksandra.gabryszak, phuc_tran.truong, arne.binder, philippe.thomas}@dfki.de
{nikola.milosevic, felix-sebastian.keese, astrid.rheinlaender}@bayer.com

Abstract

Determining the factual status of biomedical statements, whether affirmed, negated, or uncertain, is essential for accurate understanding. To support research in this area, we introduce BioRelFact, a publicly available, expert-annotated dataset of 1,767 English biomedical sentences labeled with nine relation types and five levels of epistemic commitment. Using this dataset, we evaluate eight large language models (LLMs) from the GPT, Qwen, and Gemma families for joint relation extraction and epistemic classification. Among the evaluated models, GPT-OSS-20B performs best in both tasks (F1 77.3 for relation, 65.3 for commitment), followed by GPT-4o (75.9 and 60.2), while Qwen3-8B (Thinking) shows strong performance despite its smaller size (74.6 and 57.2). Domain adaptation has mixed effects: relative to their general-purpose counterparts, MedGemma-27B improves (+3.6 F1 for relation, +4.4 for factuality), whereas Qwen2.5-Aloe-Beta-7B declines (–4.3 and –3.5, respectively). Moreover, definition-based few-shot prompts consistently yield the best results for most models, and an explorative analysis of prediction errors suggests which specific linguistic features may drive model confusions.

Keywords: relation extraction, factuality, large language models

1. Introduction

Understanding the relationships between drugs, genes, and diseases is fundamental for biomedical research and clinical decision-making. Yet, interpreting biomedical literature requires more than recognizing which relationships are mentioned: it demands an understanding of how certain those relationships are asserted. For example, the statements *"Drug X treats Disease Y"* (factual), *"Drug X may help with Disease Y"* (possible), and *"Drug X has been studied for Disease Y"* (uncommitted) all refer to a therapeutic use of a drug for a disease, but differ notably in the epistemic commitment of the sentence authors regarding the relationship. Table 1 illustrates such distinctions across common commitment levels. These nuances become especially critical in computational drug development and target discovery, where literature-derived knowledge graphs support reasoning and hypothesis generation. If speculative or uncertain statements are incorporated as factual links, such graphs may propagate misleading information, undermining their reliability for downstream analyses.

Despite its importance, epistemic nuance (also referred to as factuality) remains underrepresented in biomedical relation extraction. Most widely used datasets, such as ChemProt (Krallinger et al., 2017) and DrugProt (Miranda et al., 2021), focus on relation type identification and lack an additional factuality layer. Only a few corpora, like SemRep (Kilicoglu et al., 2017) and DiMB-RE (Hong et al., 2025), explicitly annotate semantic relations with fine-grained factuality levels.

Relation: Therapeutic_Use	
Factuality	**Example**
Factual	`Drug` treats `Disease`
Possible	`Drug` may help with `Disease`
Doubtful	`Drug` is unlikely to benefit `Disease`
Negated	`Drug` does not treat `Disease`
Uncommitted	`Drug` is being studied for `Disease`

Table 1: Example statements illustrating different levels of epistemic commitment (factuality) for the `Therapeutic_Use` relation.

Given the limitations of available datasets, large language models (LLMs) provide a promising approach to factuality-aware biomedical relation extraction, combining strong general language understanding with zero- and few-shot capabilities. However, approaches using LLMs for factuality-aware relation classification remain largely unexplored. DiMB-RE (Hong et al., 2025) evaluates GPT-4o variants in a three-factuality-label setting, but comprehensive investigations of LLMs for fine-grained, joint factuality and relation classification are still largely lacking.

To address this gap we present a dataset for

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7592–7605
11–16 May 2026.

factuality-aware biomedical relation extraction and use it to evaluate LLMs. Our main contributions are as follows:

- Introducing BioRelFact, an expert-annotated dataset of 1,767 sentences from PubMed abstracts, covering nine relation types across multiple biomedical subdomains: drug–gene, drug–disease, gene–disease, and gene variant–disease interactions, and labeled with five levels of epistemic commitment. [1]
- Performing comprehensive evaluation of eight large language models for joint relation and factuality classification, including general-purpose models (both commercial and open-weight) and biomedical models (open-weight only), using zero- and few-shot prompting strategies.
- Measuring the difficulty of classification instances, along with a linguistic analysis of the prediction errors.

2. Related Work

Biomedical Relation Extraction and Factuality Datasets. Most biomedical relation extraction datasets, such as ChemProt (Krallinger et al., 2017) and DrugProt (Miranda et al., 2021), focus on relations but largely ignore epistemic commitment. Speculative statements are usually annotated as positive relations alongside factual ones, and ChemProt's `NOT` class is sentence-level, without specifying which relation is negated. However, some corpora explicitly annotate epistemic commitment in biomedical texts. SemRep (Kilicoglu et al., 2017) labels semantic relations from PubMed abstracts with seven factuality categories: `FACT`, `PROBABLE`, `POSSIBLE`, `DOUBTFUL`, `COUNTERFACT`, `UNCOMMITTED`, and `CONDITIONAL`. DiMB-RE (Hong et al., 2025) annotates diet–microbiome associations with six levels: `FACTUAL`, `PROBABLE`, `POSSIBLE`, `DOUBTFUL`, `NEGATED`, and `UNKNOWN`. Other datasets, such as i2b2/VA (Uzuner et al., 2011), its translation (Sumait et al., 2023), Ex4CDS (Roller et al., 2022), and NegEx-Ger (Cotik et al., 2016), provide negation or factuality annotations at the entity or clinical concept level rather than for explicit inter-entity relations. BioScope (Vincze et al., 2008) provides token-level annotations for negative and speculative keywords. Thompson et al. (2011) focus on event certainty in the GENIA Event Corpus.

[1] All resources for our LLM-based classification workflow are publicly available at https://github.com/bayer-int/biomed-relation-factuality-detection.

Performance in the biomedical domain. Early domain-adapted generative models such as BioGPT (Luo et al., 2022), which was based on GPT-2 and further pre-trained on PubMed, demonstrated that generative language models are competitive for biomedical end-to-end relation extraction after fine-tuning. With the shift to GPT-3-era, billion-parameter LLMs (Brown et al., 2020), zero-/few-shot prompting became feasible. Jahan et al. (2024) find that in biomedical end-to-end relation extraction for datasets with smaller training sets, zero-shot LLMs can outperform traditional state-of-the-art models when they are only fine-tuned on the training set of these datasets. On standard RE benchmarks (ChemProt, DDI, EU-ADR and GAD), zero-/few-shot LLMs trail fine-tuned encoder models (Zhang et al., 2024; Chen et al., 2025). For example, on ChemProt, BioBERT strongly outperforms few-shot GPT-4 with 73.44 F1 versus 37.56 F1, and even fine-tuned LLaMA-2 13B, which had 46.12 F1 (Chen et al., 2025). Biomedical LLMs improve medical question answering but show mixed gains on information extraction tasks compared to general-domain counterparts (Chen et al., 2025; Garcia-Gasulla et al., 2025). Beyond LLM comparisons, prior work shows that domain-specific encoder-based transformer models outperform rule-based and classical machine learning approaches for biomedical relation extraction, while remaining robust on small and imbalanced datasets (Milošević and Thielemann, 2023). For epistemic commitment, Hong et al. (2025) report a joint evaluation that assesses whether both the relation and its factuality are predicted correctly. Their experiments are conducted in a three-label setting, collapsing the original six factuality levels. In this setting, a fine-tuned BioMedBERT model strongly outperforms zero-shot and one-shot GPT-4 variants, with LLMs tending to overpredict relatedness. Beyond LLMs, Kilicoglu et al. (2017) study factuality classification using a rule-based compositional approach based on lexical and syntactic cues versus a supervised machine learning (SVM) approach. They find that the compositional approach is more effective than the machine learning method.

3. Dataset

We introduce BioRelFact, an expert-annotated dataset of 1,767 biomedical relation instances, each consisting of a sentence containing two highlighted entities and annotated with both a relation type and a factuality label. The instances were randomly drawn from a corpus of 7,295 PubMed abstracts, covering multiple biomedical subdomains. Table 2 illustrates data examples, Section A.1 provides label definitions.

Relation	Factuality	Example
no_relation	none	*Further, [[SESN1 (Genes)]] improved [[sevoflurane (Drugs)]]-induced cell inflammation.*
drug : disease		
Causal_Effect	fact	*Paternal [[methotrexate (Drugs)]] use was associated with increased risk of [[stillbirth (Diseases)]] <..>.*
Therapeutic Use	uncommitted	*Purpose: We compared the clinical effects of <..>, [[limaprost alfadex (Drugs)]], and <..> for [[lumbar spinal stenosis (Diseases)]] <..>*
drug : gene		
Agonist	counterfact	*Expressions of ABCG2, and p-Akt but not of [[MDR-1 (Genes)]], were enhanced by NE plus [[cisplatin (Drugs)]]*
Antagonist	fact	*This could be prevented by administering <..> [[spironolactone (Drugs)]], a well-known <..> [[aldosterone receptor (Genes)]] antagonist.*
Modulates	possible	*<..> results suggested that the <..> ([[MAPK (Genes)]]) signalling pathway may be a potential mechanism of [[berberine (Drugs)]] in delaying pulmonary fibrosis.*
gene : disease		
Biomarker	counterfact	*<..> we did not verify a higher imunoexpression of [[WT1 (Genes)]] associated with [[pregnancy (Diseases)]] status.*
Causal_Effect	possible	*Conclusion: This study highlights the potential roles of specific inflammatory [[cytokines (Genes)]] in the development of [[glaucoma (Diseases)]]<..>*
Modulates	uncommitted	*Objectives: To elucidate the contribution of IL-17 family [[cytokines (Genes)]] in [[psoriasis (Diseases)]].*
geneVariant : disease		
Association	doubtful	*<..> The correlation among Glucokinase (GCK) [[rs1799884 (Gene Variant)]] polymorphism and the risk of [[gestational diabetes mellitus (Diseases)]] (GDM) remains controversial <..>*

Table 2: Examples of relations and their factuality.

3.1. Dataset Labels

Relation Types. Our relation schema was designed to capture interactions among four key biomedical entity classes: drugs, diseases, genes, and gene variants. It defines nine relation types linking these entities as listed in Table 3. Relations such as Therapeutic Use, Agonist or Biomarker exemplify the diversity of relation types included.

We assign a sentence with an annotated entity pair one of the nine relation labels if it discusses the relation assigned to the pair, irrespective of epistemic commitment (including factual, uncertain, or negated statements). We refer to such cases as *positive* instances. All other cases are assigned the label no_relation and treated as *negative instances*.

The dataset comprises 1,767 annotated sentences, of which 1,029 (58.2%) express a positive relation instance. These positive instances include 309 drug–disease, 153 drug–gene, 267 gene–disease, and 300 gene variant–disease mentions. The remaining 738 sentences (41.8%) are labeled as no_relation (drug–disease: 151, drug–gene: 291, gene–disease: 169, geneVariant–disease: 127).

Relation	Count	Percentage
no_relation	738	41.77%
drug : disease		
Causal_Effect	63	3.57%
Therapeutic_Use	246	13.92%
drug : gene		
Agonist	22	1.25%
Antagonist	89	5.04%
Modulates	42	2.38%
gene : disease		
Biomarker	52	2.94%
Causal_Effect	47	2.66%
Modulates	168	9.51%
geneVariant : disease		
Association	300	16.98%
Total	**1767**	**100.00%**

Table 3: Distribution of Relation Types

Epistemic Commitment Levels. Each sentence annotated with a positive relation type was additionally labeled with one of five epistemic commitment levels: (1) fact denotes that the sentence presents the relation as rather established or as-

Factuality level	Count	Percentage
fact	832	80.86%
uncertain		
uncommitted	121	11.76%
possible	51	4.96%
doubtful	4	0.39%
counterfact	21	2.04%
Total	**1029**	**100.00%**

Table 4: Distribution of Factuality Values for Positive Instances Only

sumed; (2) possible, as potential or probable; (3) doubtful, as unlikely; (4) counterfact, as negated; and (5) uncommitted, when the sentence gives no indication of certainty or polarity.

A closer look reveals that uncertainty rates vary across relation groups: gene variant–disease shows the highest proportion of uncertain instances (23.0%), followed by drug–disease (17.5%) and gene–disease (16.5%), while drug–gene relations show the lowest uncertainty rate at only 5.9%. Counterfacts remain rare across all groups.

3.2. Annotation Procedure and Data Quality

Given the need for strong domain expertise, we contracted two trained professionals from an external biomedical vendor to perform all annotations. Annotators followed detailed guidelines for entity recognition, relation labeling, and epistemic commitment classification, which included general instructions, label definitions and examples. In total, 2,000 sentences with relevant entity pairs were randomly selected from abstracts. Before full annotation, annotators labeled sample sentences from each relation group and received brief feedback to clarify instructions.

Entity types were initially annotated and normalized automatically using a commercial tagging system customized with Bayer-developed terminologies, and subsequently reviewed by expert annotators who flagged incorrect annotations using the label incorrect_concept_type

For each sentence containing two correctly typed entities, annotators labeled the relation discussed between them, selecting from predefined positive relation types or no_relation if no relation was mentioned. For each positive relation selected, annotators were then asked to assign exactly one epistemic commitment level.

To assess annotation reliability, we computed inter-annotator agreement (IAA) using Krippendorff's α. Agreement was very high for relation labels ($\alpha = 0.964$) and epistemic commitment levels ($\alpha = 0.963$, calculated only for sentences with a positive relation). Annotation disagreements were not retained in the dataset: sentences without full agreement on relation type, epistemic commitment level, or entity type correctness were excluded from the final release.

4. LLM-based Classification

We present a classification setup leveraging LLMs for relation and epistemic commitment classification in biomedical text. Employing an in-context learning approach, the workflow guides LLMs using modular prompts that combine task descriptions, label definitions, and annotated examples. We evaluate a range of LLMs, including general-purpose and domain-specific biomedical models of varying sizes, to compare performance and robustness. Multiple inference runs allow us to assess prediction stability.

4.1. Classification Workflow

Figure 1 shows our relation and epistemic commitment classification workflow, implemented using the LangChain framework[2]. The process begins with the specification of the input, which includes the annotation guidelines (containing label definitions and examples), a prompt template with general task instructions (for classifying relation type and epistemic commitment), label placeholders, toggles for including definitions and examples, and the target sentence with highlighted biomedical entity mentions (e.g., drugs, genes). The design of the prompt templates is modular and largely dataset-agnostic, allowing the pipeline to be reused across different corpora as long as the required input components are available.

These components are compiled into the LLM input, which consists of a structured system message that defines the task, lists valid relation types for the specific entity pair (e.g., only drug–disease relations for drug–disease pairs), optionally provides label definitions and examples, and specifies the required output format. The final user message includes the annotated input sentence and a response format specifying the JSON output schema expected from the model. The prompt is processed by a supported LLM, which returns structured output in the specified format, including the sentence ID, the predicted relation types, and their corresponding factuality values.

[2] https://github.com/langchain-ai/langchain

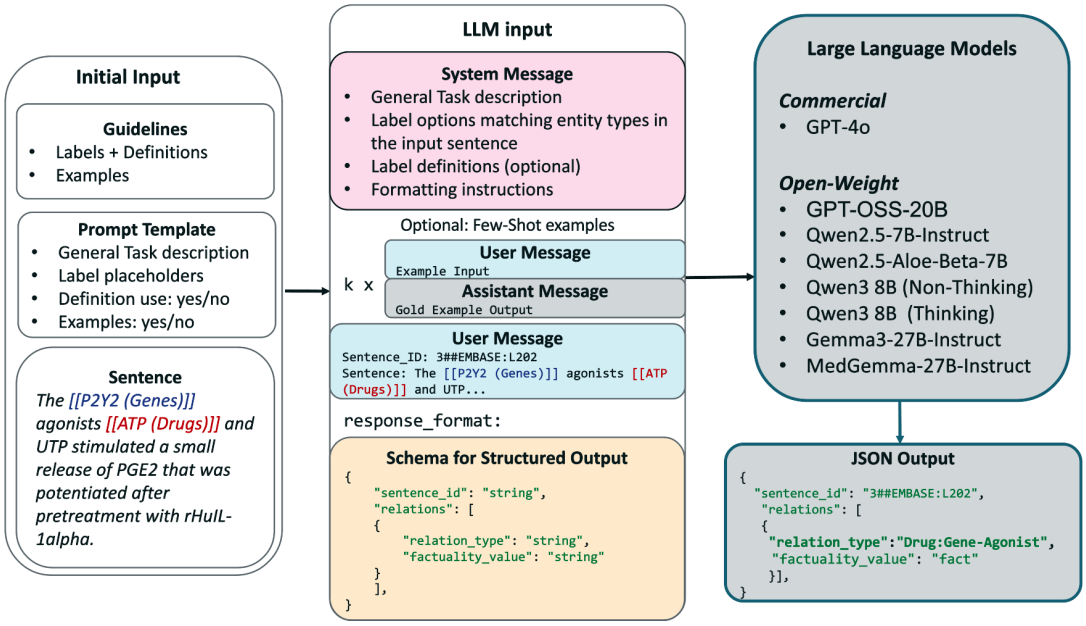

Figure 1: Overview of the LLM-based workflow for relation type and epistemic commitment classification.

4.2. Prompt Design

We evaluated prompts built from four possible components:

- General task description – a brief overview of the classification task.
- Label options – a list of relation types and factuality categories.
- Label definitions – descriptions of each relation type and factuality category.
- Few-shot examples – example sentences with labeled relations and factuality values.

The first two components (general task description and label options) are included in all prompts, while label definitions and few-shot examples are optional. Combining these components yields four prompt variants: (1) zero-shot-labels, (2) zero-shot-definitions, (3) few-shot-labels, and (4) few-shot-definitions, as summarized in Table 5. The wording of the prompts is provided in Section A.2.

At inference time, prompts are dynamically constructed based on the entity types highlighted in the sentence, presenting only compatible relation types to improve classification precision. We experiment with 2-shot and 5-shot settings, where we randomly sample 2 or up to 5 examples per relation type relevant to the annotated entity pair, along with up to the same number of negative examples. We do not explicitly control for the coverage of all factuality values within these sets.

4.3. Models

We evaluate both general-domain and biomedical-adapted LLMs across commercial APIs and open-weight model families.[3]

Prompt Variant	Contents
Zero-Shot+Labels	relation and factuality labels
Few-Shot+Labels	labels, examples
Zero-Shot+Definitions	labels and their definitions
Few-Shot+Definitions	labels, definitions, examples

Table 5: Prompt variants

As commercial baseline we used GPT-4o (Hurst et al., 2024) that offers strong language understanding capabilities. Our open-weight general-domain models include Qwen2.5-7B-Instruct (Yang et al., 2025b), Qwen3-8B (Yang et al., 2025a) (both in "Thinking" and "Non-Thinking" modes), Gemma 3-27B (Kamath et al., 2025) using the Unsloth BnB-4Bit quantized version, and GPT-OSS-20B (Agarwal et al., 2025), a reasoning model using a mixture-of-experts transformer architecture.

To assess the impact of domain adaptation, we further evaluated MedGemma-27B (Google, 2025), a medical variant of Gemma, and Qwen2.5-Aloe-Beta-7B (Garcia-Gasulla et al., 2025), a healthcare LLM built on Qwen2.5. MedGemma-27B has been exclusively trained on medical text, whereas Qwen2.5-Aloe-Beta-7B was fine-tuned on medical data with subsequent model merging and alignment (Garcia-Gasulla et al., 2025).

[3]All open-weight models were accessed via Hugging Face: Qwen/Qwen2.5-7B-Instruct, Qwen/Qwen3-8B, unsloth/gemma-3-27b-it-unsloth-bnb-4bit, openai/gpt-oss-20b, unsloth/medgemma-27b-text-it-unsloth-bnb-4bit, HPAI-BSC/Qwen2.5-Aloe-Beta-7B (https://huggingface.co).

Prompt Type	Relation Classification			Factuality Classification		
	P	R	F1	P	R	F1
GPT-4o (temperature=0.0)						
zero-shot-labels	56.3 ± 0.2	83.3 ± 0.3	67.2 ± 0.2	40.9 ± 0.5	60.6 ± 0.7	48.9 ± 0.6
2-shot-labels	64.5 ± 0.2	**85.6 ± 0.3**	73.6 ± 0.2	48.7 ± 0.1	64.6 ± 0.0	55.5 ± 0.1
zero-shot-definitions	64.9 ± 0.4	77.3 ± 0.8	70.6 ± 0.5	50.2 ± 0.2	59.7 ± 0.2	54.5 ± 0.2
2-shot-definitions	**69.6 ± 0.3**	83.3 ± 0.6	**75.9 ± 0.3**	**55.2 ± 0.6**	66.1 ± 0.5	**60.2 ± 0.5**
5-shot-definitions	65.6 ± 0.3	85.0 ± 0.3	74.0 ± 0.3	52.6 ± 0.3	**68.3 ± 0.2**	59.4 ± 0.3
GPT-OSS-20B (Reasoning Level: Low) (temperature=0.0)						
zero-shot-labels	58.9 ± 1.2	57.0 ± 2.0	57.9 ± 1.6	34.5 ± 1.3	45.8 ± 1.5	39.4 ± 1.4
2-shot-labels	66.3 ± 0.9	84.7 ± 0.6	74.4 ± 0.8	53.2 ± 0.8	68.0 ± 0.8	59.7 ± 0.8
zero-shot-definitions	66.8 ± 0.2	83.8 ± 0.4	74.4 ± 0.3	52.3 ± 0.4	67.6 ± 0.7	59.0 ± 0.5
2-shot-definitions	**69.6 ± 0.9**	**87.0 ± 0.9**	**77.3 ± 0.7**	56.9 ± 0.9	71.1 ± 0.6	63.2 ± 0.6
5-shot-definitions	68.6 ± 0.9	86.5 ± 0.7	76.5 ± 0.8	**58.5 ± 0.5**	**73.9 ± 0.3**	**65.3 ± 0.4**
Qwen2.5-7B-Instruct (temperature=0.0)						
zero-shot-labels	37.4 ± 1.3	19.5 ± 0.4	25.6 ± 0.6	9.7 ± 0.4	16.0 ± 0.6	12.1 ± 0.5
2-shot-labels	**60.7 ± 0.5**	**75.4 ± 0.8**	**67.3 ± 0.6**	45.7 ± 0.5	**57.5 ± 0.8**	50.9 ± 0.6
zero-shot-definitions	45.8 ± 0.3	55.8 ± 0.5	50.3 ± 0.4	29.0 ± 0.3	45.5 ± 0.5	35.4 ± 0.4
2-shot-definitions	58.8 ± 0.3	69.6 ± 0.3	63.8 ± 0.3	45.7 ± 0.6	54.3 ± 0.5	49.7 ± 0.5
5-shot-definitions	59.7 ± 0.2	70.1 ± 0.7	64.5 ± 0.3	**47.4 ± 0.3**	55.6 ± 0.7	**51.2 ± 0.4**
Qwen2.5-Aloe-Beta-7B (temperature=0.0)						
zero-shot-labels	40.6 ± 1.1	16.0 ± 0.3	23.0 ± 0.1	8.2 ± 0.2	13.7 ± 0.2	10.3 ± 0.2
2-shot-labels	**53.4 ± 0.3**	76.7 ± 0.1	**63.0 ± 0.2**	35.5 ± 0.0	56.8 ± 0.3	43.7 ± 0.1
zero-shot-definitions	47.4 ± 0.1	63.6 ± 0.5	54.3 ± 0.2	30.1 ± 0.2	51.4 ± 0.4	37.9 ± 0.3
2-shot-definitions	51.9 ± 0.1	**78.5 ± 0.2**	62.5 ± 0.1	37.4 ± 0.4	**59.8 ± 0.8**	46.0 ± 0.5
5-shot-definitions	51.9 ± 0.3	77.5 ± 0.1	62.1 ± 0.2	**39.8 ± 0.9**	59.6 ± 0.8	**47.7 ± 0.9**
Qwen3-8B (Non-Thinking) (temperature=0.0)						
zero-shot-labels	38.3 ± 0.1	37.8 ± 0.1	38.0 ± 0.0	17.2 ± 0.1	28.9 ± 0.2	21.6 ± 0.2
2-shot-labels	57.5 ± 0.4	**77.1 ± 0.5**	65.9 ± 0.4	44.4 ± 0.2	61.4 ± 0.3	51.5 ± 0.2
zero-shot-definitions	46.4 ± 0.1	73.6 ± 0.1	56.9 ± 0.1	36.6 ± 0.1	57.9 ± 0.1	44.8 ± 0.1
2-shot-definitions	59.1 ± 0.2	74.5 ± 0.5	65.9 ± 0.3	45.9 ± 0.4	58.1 ± 0.5	51.3 ± 0.4
5-shot-definitions	**60.1 ± 0.2**	76.1 ± 0.2	**67.2 ± 0.2**	**49.9 ± 0.3**	**63.2 ± 0.5**	**55.8 ± 0.4**
Qwen3-8B (Thinking) (temperature=0.0)						
zero-shot-labels	59.1 ± 0.7	53.5 ± 0.4	56.1 ± 0.5	29.7 ± 0.4	40.9 ± 0.1	34.4 ± 0.3
2-shot-labels	61.9 ± 0.5	84.5 ± 0.9	71.4 ± 0.7	44.5 ± 0.8	63.0 ± 0.9	52.2 ± 0.8
zero-shot-definitions	**67.2 ± 0.3**	83.4 ± 0.7	74.5 ± 0.5	**51.3 ± 0.1**	64.7 ± 0.6	**57.2 ± 0.3**
2-shot-definitions	65.9 ± 0.2	85.8 ± 0.2	74.5 ± 0.1	49.3 ± 0.3	**66.1 ± 0.2**	56.5 ± 0.2
5-shot-definitions	65.3 ± 0.2	**86.8 ± 0.5**	**74.6 ± 0.3**	46.7 ± 0.9	64.9 ± 1.1	54.3 ± 1.0
Gemma3-27B (temperature=0.0)						
zero-shot-labels	46.1 ± 0.0	58.7 ± 0.0	51.7 ± 0.0	25.4 ± 0.0	43.4 ± 0.0	32.0 ± 0.0
2-shot-labels	58.2 ± 0.2	69.4 ± 0.3	63.3 ± 0.2	47.4 ± 0.1	57.1 ± 0.1	51.8 ± 0.1
zero-shot-definitions	52.9 ± 0.0	**86.6 ± 0.0**	65.7 ± 0.0	40.2 ± 0.0	**65.8 ± 0.0**	49.9 ± 0.0
2-shot-definitions	**61.6 ± 0.2**	79.0 ± 0.1	**69.2 ± 0.1**	**50.3 ± 0.5**	64.6 ± 0.4	**56.5 ± 0.5**
5-shot-definitions	60.8 ± 0.3	78.3 ± 0.3	68.4 ± 0.2	49.7 ± 0.3	64.1 ± 0.5	56.0 ± 0.3
MedGemma-27B (temperature=0.0)						
zero-shot-labels	49.1 ± 0.1	81.0 ± 0.2	61.1 ± 0.2	38.3 ± 0.2	63.7 ± 0.2	47.8 ± 0.2
2-shot-labels	62.2 ± 0.3	74.8 ± 0.2	67.9 ± 0.3	53.8 ± 0.4	64.6 ± 0.4	58.7 ± 0.4
zero-shot-definitions	56.2 ± 2.8	**86.0 ± 0.5**	67.9 ± 1.9	43.0 ± 2.4	65.7 ± 0.1	51.9 ± 1.8
2-shot-definitions	64.3 ± 0.4	78.4 ± 0.2	70.6 ± 0.3	**55.2 ± 0.8**	67.4 ± 0.6	60.7 ± 0.7
5-shot-definitions	**64.7 ± 0.1**	83.2 ± 0.6	**72.8 ± 0.3**	54.1 ± 0.3	**69.6 ± 0.7**	**60.9 ± 0.4**

Table 6: Relation & Factuality classification scores (Precision, Recall, F1) across prompt types and models. **Bold** = best within model (as in source tables), **bold+underlined** = best overall.

Model	Drug–Disease	Drug–Gene	Gene–Disease	Variant–Disease	Micro Avg
GPT-4o	86.0	63.1	62.9	84.0	75.9
GPT-OSS-20B	**86.9**	**71.1**	**62.9**	**86.7**	**77.3**
Qwen2.5-7B-Instruct	74.4	25.6	58.0	82.9	67.3
Qwen2.5-Aloe-Beta-7B	67.6	26.4	54.6	81.9	63.0
Qwen3 8B Non-Thinking	77.1	41.0	54.6	83.1	67.2
Qwen3 8B Thinking	85.3	62.8	62.2	83.9	74.6
Gemma3-27B-Instruct	80.4	52.4	51.4	83.3	69.2
MedGemma-27B-Instruct	82.8	57.0	59.9	83.0	72.8

Table 7: Weighted F1 per relation group across best settings for models. **Bold** = best model.

Model	Fact	Possible	Doubtful	Uncommitted	Counterfact	Micro Avg
GPT-4o	67.3	34.2	**49.2**	46.0	54.0	60.2
GPT-OSS-20B	**71.3**	37.3	34.0	**54.8**	50.4	**65.3**
Qwen2.5-7B-Instruct	56.9	34.5	13.5	39.6	24.9	51.2
Qwen2.5-Aloe-Beta-7B	55.5	18.9	0.0	22.2	19.4	47.7
Qwen3 8B (Non-Thinking)	61.4	29.1	5.9	44.6	42.3	55.8
Qwen3 8B (Thinking)	67.2	26.2	0.0	19.7	19.5	57.2
Gemma3-27B-Instruct	61.7	35.0	22.2	45.4	**61.4**	56.5
MedGemma-27B-Instruct	64.6	**46.3**	32.1	50.5	52.8	60.9

Table 8: Factuality classification F1 scores per class across best setting for models. **Bold** = best model.

5. Experiments and Results

We evaluate the models from Section 4.3, using the settings shown in the result table headers. Each configuration is run three times, and we report the mean and standard deviation ($\pm$) across runs.

5.1. Relation Classification

The results for relation classification are presented in Table 6. GPT-OSS-20B achieved the highest overall F1 of 77.3 with a 2-shot definitions prompt, followed closely by its 5-shot definitions result at 76.5. GPT-4o (2-shot definitions, 75.9) and Qwen3-8B (Thinking, 5-shot definitions, 74.6) performed competitively, with Qwen3-8B showing strong results despite its smaller size. GPT-OSS-20B also attained the highest recall of 87.0 and, together with GPT-4o, the highest precision of 69.6. Reasoning-enabled models (GPT-OSS-20B and Qwen3-8B-Thinking) achieve these results at the cost of increased token usage and slower inference due to longer outputs.

Domain adaptation effects are mixed. Comparing best settings, MedGemma-27B improves over its general-domain counterpart Gemma3-27B across F1 (72.8 vs. 69.2), precision (64.7 vs. 61.6), and recall (83.2 vs. 79.0). In contrast, within the Qwen2.5 family, the medical variant Qwen2.5-Aloe-Beta-7B underperforms Qwen2.5-7B-Instruct in F1 (63.0 vs. 67.3) and precision (53.4 vs. 60.7), while achieving a slightly higher recall (76.7 vs. 75.4).

Next we analyzed extending label prompts with few-shot examples versus definitions. Qwen2.5 models and Qwen3-8B Non-Thinking gain more from examples, while Qwen3-8B Thinking, Gemma, and GPT families benefit similarly from either. For example, Qwen2.5-7B-Instruct improves from F1 25.6 $\rightarrow$ 50.3 $\rightarrow$ 67.3 (0-shot labels $\rightarrow$ 0-shot definitions $\rightarrow$ 2-shot labels), whereas GPT-4o goes from 67.2 $\rightarrow$ 70.6 $\rightarrow$ 73.6 respectively. Increasing from zero- to two-shot consistently improves performance, with largest gains for labels and moderate gains for definitions. Further increasing to five shots has a small decreasing or increasing impact. Overall, combining definitions with examples yields the strongest performance for most models, except Qwen2.5, where two-shot labels alone yield the highest F1.

Finally, we compared performance across relation groups for each model's best prompting setting (Table 7). Drug–Disease relations are the easiest to classify, with F1 scores ranging from 67.6 (Qwen2.5-Aloe-Beta-7B) to 86.9 (GPT-OSS-20B), indicating consistently strong performance across models. Drug–Gene relations are more challenging, particularly for smaller models (F1 as low as 25.6), while larger models reach up to 71.1. Gene–Disease relations also remain difficult, with F1 spanning 51.4–62.9 across models. In contrast, GeneVariant–Disease/Phenotype relations are relatively easy across models (F1 81.9–86.7), likely due to having only one general category. Across models' best settings, the hardest relations are `gene:disease-Causal_Effect` (up to F1 43.5) and `drug:gene-Modulates` (F1 48.5), whereas the easiest is `drug:disease-Therapeutic_Use` (F1 87.8).

5.2. Factuality Classification

Results for factuality classification are presented in Table 6. We adopt strict factuality evaluation, where a prediction is considered correct only if both the relation type and its factuality are correctly predicted. GPT-OSS-20B is the top-performing model overall, achieving its highest F1 (65.3), precision (58.5), and recall (73.9) with the 5-shot definitions prompt, while the 2-shot definitions prompt yields the second-best results across all metrics. MedGemma-27B follows as the second-best model (F1 60.9), ahead of GPT-4o (F1 60.2) and Gemma3-27B (F1 56.5). Within the Qwen3-8B family, the Thinking variant outperforms its Non-Thinking counterpart (F1 57.2 vs. 55.8), although it is less competitive with the top models than in relation extraction.

Regarding domain adaptation, trends similar to relation classification are observed. Comparing best settings, MedGemma-27B outperforms its general-domain counterpart Gemma3-27B across all factuality metrics: F1 (60.9 vs. 56.5), precision (54.1 vs. 50.3), and recall (69.6 vs. 64.6). Within the Qwen2.5 family, Qwen2.5-7B-Instruct surpasses the medical variant Qwen2.5-Aloe-Beta-7B in F1 (51.2 vs. 47.7) and precision (47.4 vs. 39.8), whereas the latter attains higher recall (55.6 vs. 59.6).

For prompt design, few-shot label variant generally improves over zero-shot definitions, with notable gains for Qwen2.5-7B-Instruct (F1 35.4 $\rightarrow$ 50.9), Qwen3-8B Non-Thinking (44.8 $\rightarrow$ 51.5), and MedGemma-27B (52.9 $\rightarrow$ 58.7). Few-shot definitions further boost performance for most models, such as GPT-4o (0-shot definitions 54.5 $\rightarrow$ 2-shot definitions 60.2) and GPT-OSS-20B (59.0 $\rightarrow$ 63.2), although improvements from 2- to 5-shot definitions are generally smaller and occasionally slightly negative (e.g., Qwen3-8B Thinking: 2-shot 56.5 $\rightarrow$ 5-shot 54.3).

Finally, we analyzed performance across factuality classes for best models settings (Table 8). Factual statements are the easiest to classify, with F1 scores ranging from 55.5–71.3 across best settings for models. Rare classes remain challenging, and performance estimates are less reliable due to limited sample sizes: counterfact F1 spans 19.4–61.4, doubtful 0.0–49.2, possible 18.9–46.3, and uncommitted 19.7–54.8. Smaller models generally underperform on these rare classes; however, for the selected best configuration, Qwen3-8B Non-Thinking achieves notably higher F1 than the Thinking model for `counterfact` (42.3 vs. 19.5) and `uncommitted` (44.6 vs. 19.7), although the Thinking variant outperforms on rare classes in other prompt settings.

5.3. Error Analysis

This analysis builds on (Alt et al., 2020), who examined data difficulty and linguistically categorized false predictions in relation classification.

Measuring Prediction Difficulty. To assess prediction difficulty, we grouped sentences by the proportion of correct test runs for relations, factuality, and their combination. Model behavior was analyzed across 120 runs per sentence (8 models $\times$ 5 prompts $\times$ 3 runs). Using threshold cutoffs, we defined three difficulty levels: easy ($> 70\%$ correct), medium (30–70%), and challenging ($< 30\%$). Table 9 shows the distribution of categories. Overall,

	Easy	Medium	Hard	Total
	$> 70\%$	30–70%	$< 30\%$	
Relation	832	501	434	1767
Factuality	742	554	471	1767
Combined	560	622	585	1767

Table 9: Distribution of sentences by difficulty level.

relation predictions are slightly easier than factuality predictions, with 832 sentences (47.1%) classified as easy for relations versus 742 (42.0%) for factuality. When considering both criteria jointly, easy cases drop to 560 (31.7%) and challenging ones rise to 585 (33.1%), indicating that factuality errors substantially drive overall difficulty. Medium-difficulty sentences remain balanced across categories, with a modest increase in the combined measure (622, 35.2%), suggesting that many predictions are only partially correct.

Factuality Classification Errors. In an explorative analysis, we examined a sample of false factuality predictions to identify sentence features that may cause model confusion and lead to incorrect labels. Table 10 illustrates examples of false factuality predictions. The most frequent issue is the presence of a lexical distractor, i.e., an extra marker whose scope does not pertain to the existence of the target relation (e.g., a negated relation with *not*, alongside another statement containing *might* that indicates possibility). We also observed that some factuality triggers are systematically misclassified; for example, uncommitted relations hedged by *to hypothesize* were labeled as possible. This may result from ambiguous interpretation: in scientific contexts it signals an uncommitted relations, but it is also close in meaning to possibility. Moreover, contrastive negations hedged with *but not*, where the same factual relation holds but not between the highlighted entities, are often misclassified as facts. Finally, models such as Gemma, MedGemma, and

sentence	gold labels	prediction
We hypothesized that [[SIRT1 (Genes)]] activator [[resveratrol (Drugs)]] alleviates LBP and anxiety via promotion of osteogenesis in the porous endplates.	`Agonist fact`	`Agonist possible`
Expressions of ABCG2, and p-Akt but not of [[MDR-1 (Genes)]], were enhanced by NE plus [[cisplatin (Drugs)]] when compared to cisplatin only in both cell lines.	`Agonist counterfact`	`Agonist fact`
Since this [[p.Pro82Leu (Gene Variant)]] variant was not found in the [[psoriasis vulgaris (Diseases)]] and control groups in their study, they speculated that this variant might lead to exacerbated inflammatory responses.	`Association counterfact`	`Association fact`
We hypothesize that [[ibrutinib (Drugs)]] has a direct antitumor effect in melanoma cell lines and that treatment of [[metastatic melanomas (Diseases)]] with ibrutinib induces antitumor responses.	`Therapeutic_Use uncommitted`	`Therapeutic_Use possible`

Table 10: Examples of false factuality predictions

Qwen occasionally predicted a factuality label for no-relation cases, despite the instruction to assign a factuality label only to positive cases and the few-shot examples using none in such instances.

6. Conclusions

In this work, we present BioRelFact, a publicly available, expert-annotated dataset for joint biomedical relation extraction and factuality classification, and benchmark eight large language models. GPT-OSS-20B achieves the highest overall performance, while the smaller Qwen3-8B (Thinking) remains competitive. Domain adaptation shows mixed effects: MedGemma improves over Gemma, whereas Qwen2.5-Aloe underperforms relative to its general-domain counterpart. Definition-based prompts generally outperform label prompts, with 2-shot prompting offering the best trade-off. These results highlight the importance of careful prompt design and domain-aware adaptation for effective biomedical information extraction. In future work, we aim to automatically expand the dataset and improve difficulty assessment. A refined model ensemble-based framework could support these efforts by improving reliability estimates, detecting hard cases, and facilitating semi-automatic annotation of synthetic data.

7. Acknowledgements

We thank the anonymous reviewers for their insightful feedback. We are also grateful to our domain experts for tirelessly reviewing and annotating the data set presented here, and to our students — P. Danilovskaia, E. Kara and E. Eberle — for their support in the linguistic analysis. This research was funded by the Federal Ministry of Education and Research (BMBF) within the TRAILS project (Grant No. 01IW24005).

8. Limitations

While our study provides insights into LLM performance for the joint classification of relation and factuality, several limitations remain.

First, our dataset has several inherent constraints. Automatic sentence segmentation and pre-annotation of entity types improve efficiency but may introduce noise and biases from the underlying systems. Moreover, some relation and factuality classes are underrepresented, which may reduce the reliability of performance estimates for rare categories. Additionally, relation and factuality labels are assigned at the sentence level, which may not capture broader discourse context.

Second, our evaluation is limited to this dataset. While similar noise and label imbalances exist in other biomedical resources, testing on additional datasets would help assess the generalizability of our findings.

Third, we do not include a supervised encoder baseline (e.g., BioBERT) trained on the annotated dataset. Given the relatively small size and class imbalance of the dataset, particularly for rare factuality categories, fully supervised training may be challenging. Expanding the dataset to increase its size and improve class balance could facilitate more robust supervised encoder training and represents a valuable direction for future work.

Fourth, our difficulty measure has several constraints: it depends on the choices made for models, prompts, and thresholds. Although we mitigate this by selecting a reasonable and diverse set of models, prompts, and threshold values, the threshold-based categorization remains approximate.

Finally, a more detailed analysis of errors, including the correlation between instance difficulty and its features, would provide a deeper understanding of model performance and the factors contributing to model failures.

9. Ethics

Our study raises several ethical considerations regarding dataset creation, annotation, and model use.

Data sourcing. All sentences in our dataset were drawn from publicly available PubMed abstracts, which are accessible under open licenses. No patient records, clinical notes, or other sensitive personal health data were included. As such, the dataset does not contain personally identifiable information (PII). Nevertheless, biomedical literature may contain potentially sensitive associations (e.g., relating genes, diseases, or therapies).

Annotation. Annotations were performed by trained biomedical professionals contracted from a third-party vendor. Annotators were compensated at fair market rates, and were provided with detailed guidelines to ensure consistency.

Intended use and limitations. The dataset and models were developed and evaluated for research on biomedical information extraction in scientific writing. They are not designed for direct clinical use or medical decision-making. Misinterpretation or overreliance on automatically extracted relations could pose risks if applied in medical contexts. The resource will be released under the BSD 3-Clause License, which permits broad reuse, including commercial applications, while providing an "as is" warranty disclaimer. Users are responsible for ensuring appropriate use and compliance with applicable regulations.

10. Bibliographical References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

BioCreative VI. 2017. Annotation manual of CHEMPROT interactions between chemical entity mentions (cems) and gene and protein related objects (gpros). Technical report, BioCreative Challenge. Version 6.0, August 1, 2017.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, et al. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1):3280.

Viviana Cotik, Roland Roller, Feiyu Xu, Hans Uszkoreit, Klemens Budde, and Danilo Schmidt. 2016. Negation detection in clinical reports written in german. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining, BioTxtM@COLING 2016, Osaka, Japan, December 12, 2016*, pages 115–124. The COLING 2016 Organizing Committee.

Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, et al. 2025. The aloe family recipe for open and specialized healthcare llms. *arXiv preprint arXiv:2505.04388*.

Google. 2025. MedGemma Hugging Face. https://huggingface.co/collections/google/medgemma-release-680aade845f90bec6a3f60c4. Accessed: 2025-07-13.

Gibong Hong, Veronica Hindle, Nadine M Veasley, Hannah D Holscher, and Halil Kilicoglu. 2025. Dimb-re: mining the scientific literature for diet-microbiome associations. *Journal of the American Medical Informatics Association*, 32(6):998–1006.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, 171:108189.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Halil Kilicoglu, Graciela Rosemblat, and Thomas C. Rindflesch. 2017. Assigning factuality values to semantic relations extracted from biomedical research literature. *PLOS ONE*, 12(7):1–20.

Martin Krallinger, Obdulia Rabal, Saber Ahmad Akhondi, Martín Pérez Pérez, Jesus Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio Baso López, Umesh K. Nandal, Erin M. van Buel, Ambika Chandrasekhar, Marleen Rodenburg, Astrid Lægreid, Marius A. Doornenbal, Julen Oyarzábal, Anália Lourenço, and Alfonso Valencia. 2017. Overview of the biocreative vi chemical-protein interaction track. *Proceedings of the BioCreative VI Workshop,*, 141-146.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Nikola Milošević and Wolfgang Thielemann. 2023. Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *Journal of Web Semantics*, 75:100756.

Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2021. Overview of drugprot biocreative vii track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*. 2023 Journal version: https://doi.org/10.1093/database/baad080.

Roland Roller, Aljoscha Burchardt, Nils Feldhus, Laura Seiffe, Klemens Budde, Simon Ronicke, and Bilgin Osmanodja. 2022. An annotated corpus of textual explanations for clinical decision support. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2317–2326, Marseille, France. European Language Resources Association.

Mohammed Bin Sumait, Aleksandra Gabryszak, Leonhard Hennig, and Roland Roller. 2023. Factuality detection using machine translation - a use case for german clinical text. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), September 19-21, 2023, Ingolstadt, Germany*, pages 85–92. Association for Computational Linguistics.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2011. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12(1):393.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, et al. 2025b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jeffrey Zhang, Maxwell Wibert, Huixue Zhou, Xueqing Peng, Qingyu Chen, Vipina K Keloth, Yan Hu, Rui Zhang, Hua Xu, and Kalpana Raja. 2024. A study of biomedical relation extraction using GPT models. *AMIA Summits on Translational Science Proceedings*, 2024:391–400.

A. Appendix A

A.1. Dataset and Annotation Guidelines

The complete annotation guidelines are part of the publicly available repository associated with this paper. For a quick overview we present here the label definitions for all entity types[4], factuality and relation labels, which we provided to the annotators and used in the prompts where it was applicable (Tables 11, 12, 13).

Our guidelines are partially based on the ChemProt guidelines (BioCreative VI, 2017); however, for some of their defined relations, we adopted broader definitions. Moreover, although 11 positive relation types were available during annotation, only 9 are represented in the final dataset, as no instances of two relation types occurred in the sampled sentences (`Contraindication` and `Inverse Agonist`).

A.2. Prompts

In Section 4.2 we described the prompt design. Here we present the complete system-message template used in our experiments. The template in the Figure 2 illustrates the maximal prompt configuration, including placeholders for dynamically inserted entity types, relation and factuality labels, and optionally label definitions.

The placeholders such as `{entity_types}`, `{relation_names}` are instantiated dynamically at inference time based on the highlighted entity types in the input sentence. For example, if the highlighted entities correspond to a drug–disease pair, the set of candidate relation labels is restricted to those compatible with that type combination (e.g., `Causal_Effect`, `Therapeutic_Use`, and `no_relation`). The placeholder `{facticity_names}` is always filled with all possible factuality values.[5] Moreover, in definition-based prompt variants, the placeholders `{relation_definitions}` and `{facticity_definitions}` are populated with relevant textual descriptions of label values (Section A.1); in label-only variants, these sections are omitted.

[4]To ensure consistency with the relation type labels during manual annotation and LLM experiments, entities originally pre-labeled as `Indications` and `Human Genes (PH)` were mapped to `Diseases` and `Genes` respectively. The latter mapping also aimed to simplify the verification process, as the annotation task primarily involved confirming whether an entity was a gene, rather than distinguishing among finer biological subcategories.

[5]The implementation and repository use the term *facticity* for the epistemic commitment categories referred to as *factuality* in the paper. The label definitions and experimental setup are otherwise identical.

entity type	definition
`Drug`	'real' drugs, chemical compounds (including Leadmine compounds) and hormones
`Disease`	diseases, symptoms, conditions, and disease-related phenotypes
`Gene`	genes, proteins, mRNA and other gene products
`Gene Variant`	genomic/protein variants (including substitutions, deletions, insertions, and others)

Table 11: Entity types: definitions

factuality type	definition
`fact`	a relation is assumed or presented as a known, established fact.
`counterfact`	a relation is assumed or presented as rather a false or negated fact.
`possible`	a relation is assumed or presented as a possible, probable, or potential association.
`doubtful`	a relation is assumed or presented as a doubtful association.
`uncommitted`	a relation is discussed; however, it is not clear whether it is factual, counterfactual, possible or doubtful, often due to insufficient evidence or information (e.g., because the sentence states that the relation is to be investigated first or because a study did not allow a conclusion about the relation state).
`null`	no relation is discussed; therefore, the factuality value cannot be specified.

Table 12: Factuality types: definitions and examples

The prompt further specifies a structured response format: model outputs must follow a predefined JSON schema containing a `sentence_id`, a predicted `relation_type`, and a `facticity_value`. While our code supports batched requests, each request contained a single sentence in our experiments. After generation, each model response is parsed as JSON and validated against the schema, which enforces well-formedness, required field presence, correct hierarchical structure, and datatype consistency. However, the schema enforces only structural and type-level constraints; membership in the predefined closed set of relation and factuality labels is specified in the prompt but not encoded as enumerated constraints in the validation schema, which may lead to lexical deviations despite otherwise structurally valid outputs. In our experiments, such deviations occurred almost exclusively for relation

Relation	Definition
drug : disease	
Therapeutic_Use	This label refers to the discussion of indications of drugs for treating, preventing or decreasing a disease, or to relieve disease symptoms.
Contraindication	This label refers to the discussion of a disease (or a medical condition in general) as a reason to not receive a specific drug treatment (e.g. due to potential worsening of the condition itself or other severe harm. If the potential harm is not mentioned, then the phrasing must clearly imply a contraindication (e.g. phrases like "drug must not be used", "is contraindicated", etc.). Do not select this label if the sentence only mentions the necessity of taking pre-cautions when taking the medicine while having a specific medical condition.
Causal_Effect	This label refers to the discussion of a drug causing a disease or negative medical conditions (including adverse effects).
gene : disease	
Biomarker	This label refers to a specific gene or a set of genes whose presence, expression, or mutation is associated with the occurrence, progression, or risk of a particular disease.
Modulates	This label refers to the discussion whether a gene is responsible for preventing, decreasing or increasing, alleviating or worsening a disease. Do not select this label if the sentence only discusses similar topics, such as genes causing a disease, gene being a biomarker or a target for the given disease with no more specific information on the discussed association type with the disease.
Causal_Effect	This label refers to the discussion whether activation, mutation or inhibition, or any other action over a gene is causing a given disease. Do not select this label if the sentence only discusses similar topics, such as genes being a modulator, a biomarker or a target for the given disease with no more specific information on the discussed association type with the disease.
drug : gene	
Agonist	This label refers to the discussion whether a drug activates or increases gene activity. It might happen directly or indirectly, binding to a gene or its receptor is not a necessary condition to select this label. Do not select this label if the sentence discusses inverse agonism, for which there is a separate label.
Inverse_Agonist	This label refers to discussing inverse agonists, i.e. a special case of agonist, which activates and decreases gene activity.
Antagonist	This label refers to the discussion whether a drug blocks or decreases gene activity. It might happen directly or indirectly, binding to a gene or its receptor is not a necessary condition to select this label.
Modulates	This label refers to the discussion of drugs as allosteric modulators, compounds that increase or decrease the action of an (primary or orthosteric) agonist or antagonist by combining with a distinct (allosteric or allotropic) site on the receptor macromolecule. Also select this label if a drug regulates genes, but no information is available on whether the drug activates, increases, blocks or decreases gene activity.
geneVariant : disease	
Association	This label refers to the discussion of any association stated between the highlighted gen variant and a disease (including disease related phenotypes).
NO_RELATION	Select NO_RELATION if the concept types of highlighted phrases are correct, but no discussion of a listed relation between those concepts can be inferred by an expert. If a relation is described in very general, vague terms, so that a no relation is either explicitly discussed or its implication is clear for an expert, select the label 'NO_RELATION'.

Table 13: Relation labels grouped by entity-type pairs.

labels in open-weight models, predominantly under zero-shot prompting, with few-shot prompting — especially combined with definitions — consistently reducing rates across models to near-zero.

In few-shot prompt variants, we prepend k labeled demonstrations as separate user/assistant message pairs before the target input. Each demonstration consists of an example sentence with highlighted concepts (user message) and the corresponding gold relation label and factuality value (assistant message). These demonstrations use the same input and output format as the target instance; they are not embedded within the system-message template itself.

```
Your task is to analyze input sentences containing highlighted concept pairs that
correspond to {entity\_types}. For each sentence, answer two questions about the
relationship expressed between the highlighted concepts:
1. **Question 1**: Does the sentence discuss any of the listed relations involving
the highlighted concepts?

2. **Question 2**: If the sentence discusses a relationship involving the
highlighted concepts, is the relation presented as an established fact, a
counterfact, a possibility or doubt, or an uncommitted association?

Possible answers:
- **Question 1**: {relation_names}
- **Question 2**: {facticity_values}

### **Relation Name Structure**
Relation names are structured as `Entity_Type1:Entity_Type2 - Relation`.
The relationship direction is always from `Entity_Type1` (Head) to `Entity_Type2`
(Tail).

**Examples**:
- `Chemical : Gene - Agonist` means "Chemical (Head) acts as an agonist for Gene
(Tail)."
- `Chemical : Disease - Therapeutic_Use` means "Chemical (Head) has a therapeutic
use for Disease (Tail)."

### **Relation Definitions**
Refer to the following relation definitions when choosing your answer to **Question
1**:
**{relation_definitions}**

### **Facticity Definitions**
Refer to the following facticity definitions when choosing your answer to
**Question 2**:
**{facticity_definitions}**

### **Important Constraints**
The `sentence_id` must match **EXACTLY** as provided. Do not modify it.
You must choose answers only from the allowed answer list.

The output should be formatted as a JSON instance that conforms to the JSON schema
below.

Here is the output schema:

{
  "sentence_id": "string",
  "relations": [
    {
      "relation_type": "string",
      "facticity_value": "string | null"
    }
  ]
}
```

Figure 2: System prompt template for relation type and factuality classification. Placeholders in curly braces are filled dynamically per relation schema.