From Facts to Hypotheses:

Joint Detection of Biomedical Relations and Epistemic Commitment

Aleksandra Gabryszak*, Phuc Tran Truong*, Arne Binder*, Nikola Milošević†, Felix-Sebastian Keese†, Astrid Rheinländer†, Philippe Thomas*
*German Research Center for Artificial Intelligence (DFKI) | †Bayer AG, Berlin

Task

Joint Detection of Relation & Epistemic Commitment

Given a statement in a sentence with two highlighted entities, predict:

Relation: Which relation is discussed?

Epistemic Commitment: How committed is the statement author to the factual state of the relation?

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`

Motivation

Factuality-aware extraction of biomedical relations is crucial for computational drug development

- Standard relation extraction alone has limitations

  → it may either propagate misleading information (if speculative statements are treated as factual)
  → or lose important nuances (if uncertain or negated relations are discarded).

Research Gap & Contributions

- Most RE datasets ignore factuality, e.g., collapsing speculative and factual relations or treating negated relations as unmentioned

  → we publish dataset BioRelFact labeled with relations and epistemic commitment

- LLMs for factuality-aware biomedical RE are largely unexplored.

  → we test 8 LLMs (commercial vs. open weight, general purpose vs. biomedical)

BioRelFact Dataset

Creation of dataset for relation and epistemic commitment detection

- 1,767 sentences from PubMed abstracts
- Annotated by biomedical experts

Relation	Count
no_relation	738
drug : disease	
Causal_Effect	63
Therapeutic_Use	246
drug : gene	
Agonist	22
Antagonist	89
Modulates	42
gene : disease	
Biomarker	52
Causal_Effect	47
Modulates	168
geneVariant : disease	
Association	300
Total	1767

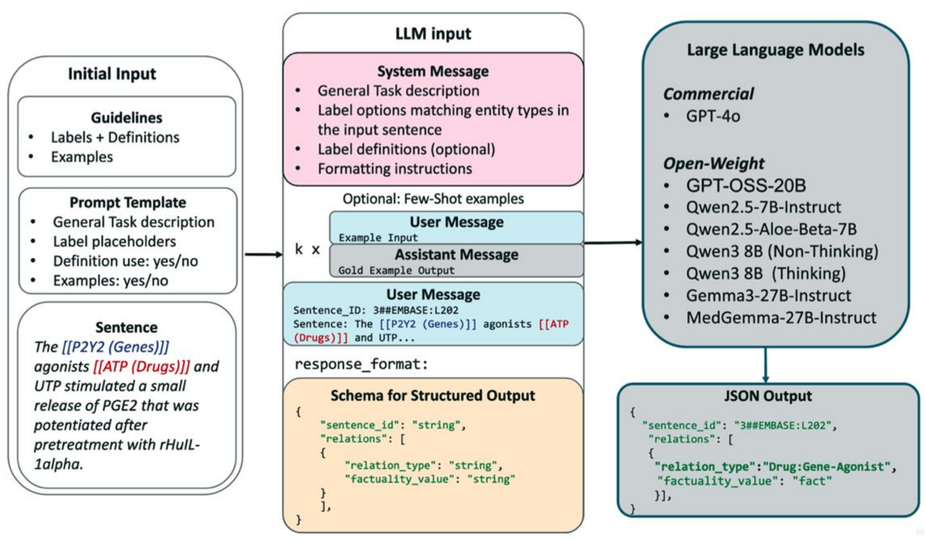
Factuality level	Count
fact	832
uncertain	
uncommitted	121
possible	51
doubtful	4
counterfact	21
Total	1029

LLM-based Classification

Prompt Variants

- **Zero-Shot + Labels** = relation and factuality labels only
- **Zero-Shot + Definitions** = labels with textual definitions
- **Few-Shot + Labels** = labels with 2 or 5 examples
- **Few-Shot + Definitions** = labels, definitions, examples

Classification Workflow



Relation Classification

Best Settings for each Model

Model	Prompt	P	R	F1
GPT-OSS-20B	2-shot def.	**69.6**	**87.0**	**77.3**
GPT-4o	2-shot def.	**69.6**	83.3	75.9
Qwen3-8B (T)	5-shot def.	65.3	86.8	74.6
MedGemma-27B	5-shot def.	64.7	83.2	72.8
Gemma3-27B	2-shot def.	61.6	79.0	69.2
Qwen3-8B (NT)	5-shot def.	60.1	76.1	67.2
Qwen2.5-7B	2-shot lab.	60.7	75.4	67.3
Qwen2.5-Aloe	2-shot lab.	53.4	76.7	63.0

F1 by Relation Group (5 best models)

Model	Drug-Disease	Drug-Gene	Gene-Disease	Variant-Disease
GPT-OSS-20B	**86.9**	**71.1**	**62.9**	**86.7**
GPT-4o	86.0	63.1	**62.9**	84.0
Qwen3-8B (T)	85.3	62.8	62.2	83.9
MedGemma-27B	82.8	57.0	59.9	83.0
Gemma3-27B	80.4	52.4	51.4	83.3

Factuality Detection

Best Settings for each Model

Model	Prompt	P	R	F1
GPT-OSS-20B	5-shot def.	**58.5**	**73.9**	**65.3**
MedGemma-27B	5-shot def.	54.1	69.6	60.9
GPT-4o	2-shot def.	55.2	66.1	60.2
Qwen3-8B (T)	0-shot def.	51.3	64.7	57.2
Gemma3-27B	2-shot def.	50.3	64.6	56.5
Qwen3-8B (NT)	5-shot def.	49.9	63.2	55.8
Qwen2.5-7B	5-shot def.	47.4	55.6	51.2
Qwen2.5-Aloe	5-shot def.	39.8	59.6	47.7

(Strict factuality evaluation: factuality prediction counted as correct if relation is correct)

F1 by Factuality value (5 best models)

Model	Fact	Poss.	Doubt.	Uncom.	Counterf.
GPT-OSS-20B	**71.3**	37.3	34.0	**54.8**	50.4
MedGemma-27B	64.6	**46.3**	32.1	50.5	52.8
GPT-4o	67.3	34.2	**49.2**	46.0	54.0
Qwen3-8B (T)	67.2	26.2	0.0	19.7	19.5
Gemma3-27B	61.7	35.0	22.2	45.4	**61.4**

Error Analysis

Difficulty Analysis

- sentences grouped by the proportion of correct test runs for relations, factuality, and their combination.

	Easy (>70%)	Medium	Hard (<30%)
Relation	832 (47%)	501 (28%)	434 (25%)
Factuality	742 (42%)	554 (31%)	471 (27%)
Combined	560 (32%)	622 (35%)	585 (33%)

(Difficulty across 120 runs/sentence
(8 models × 5 prompts × 3 runs)

Common Error Patterns:

- **Lexical factuality distractors**

  "*... GENE activator DRUG may alleviate anxiety*"
  → prediction: Agonist, possible instead of fact

- **Ambiguity of lexical cues**

  "*... we hypothesize that DRUG reduces DISEASE*"
  → prediction: Therapeutic Use, possible instead of uncommitted

- **Contrastive negation**:

  "*Expressions of ... but not of GENE were enhanced by DRUG*"
  → prediction: Agonist, fact instead of counterfact

Contributions & Key Findings

- **Dataset BioRelFact** annotated with relations and epistemic commitment
- **Evaluation of 8 LLMs**
  - Best Model: GPT-OSS-20B
  - Qwen3-8B (Thinking) a strong competitor despite smaller size
- **Prompt design matters:**
  - definitions + few-shot perform best;
  - 2-shot often provides the best trade-off between performance and efficiency
- **Domain Adaptation shows mixed results:**
  - ✓ MedGemma-27B > Gemma3:
  - × Qwen2.5-Aloe < Qwen2.5:
- **Uncertain and negated relations are challenging**

Resources (available at conference start): https://github.com/Bayer-Group/biomed-relation-factuality-detection

Funded by German Federal Ministry of Research, Technology and Space (BMFTR) as part of the TRAILS Project (Grant No. 01IW24005)