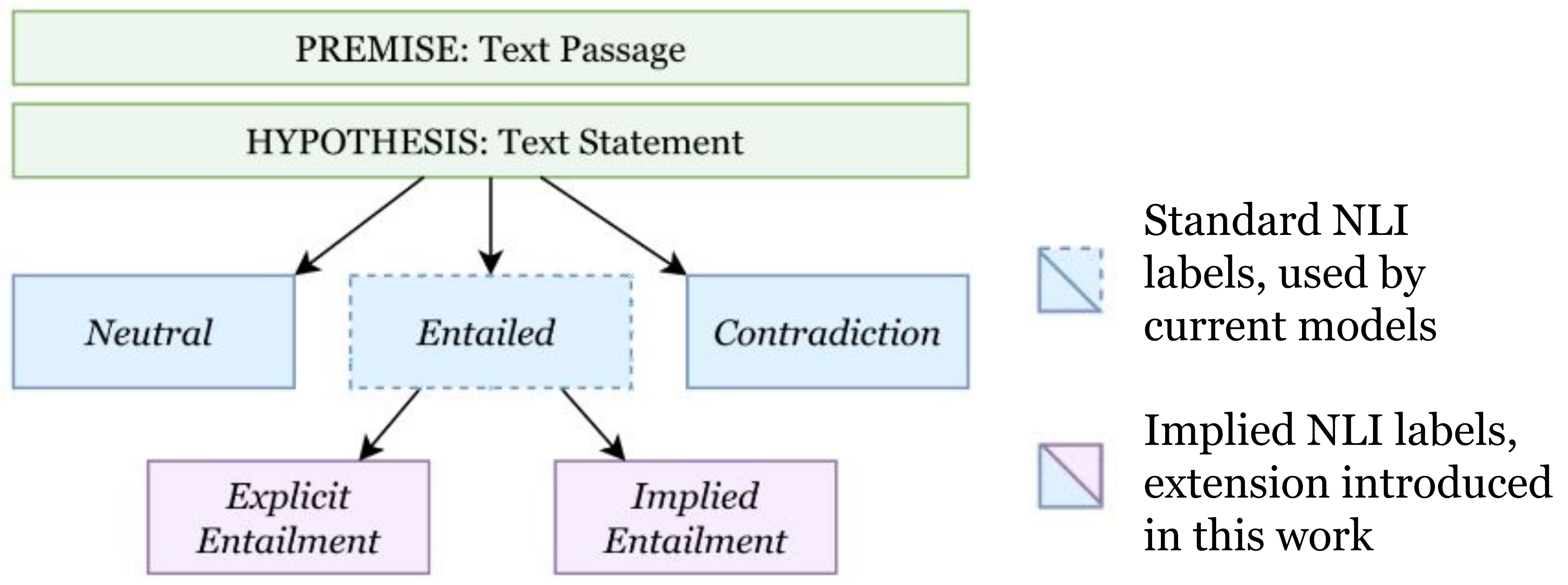
Google DeepMind

Entailed Between the Lines: Incorporating Implication into NLI

Shreya Havaldar, Hamidreza Alvari, John Palowitch,
Mohammad Javad Hosseini, Senaka Buthpitiya, Alex Fabrikant

ACL 2025 VIENNA
JULY 27 - AUGUST 1

Explicit vs. Implied Entailment


PREMISE: Text Passage

HYPOTHESIS: Text Statement

Neutral *Entailed* *Contradiction*

Explicit Entailment *Implied Entailment*

▨ Standard NLI labels, used by current models

▨ Implied NLI labels, extension introduced in this work

Explicit entailment: Follows directly from text's lexical semantics and syntax. *Synonymy, Paraphrasing, Pronominal co-reference, Bridging, Other endophora*

Implied entailment: Requires an additional cognitive deduction. *Logical reasoning, World knowledge, Conversational pragmatics, Figurative language*

Current NLI Datasets are Insufficient

Motivating RQ1: Do current NLI datasets contain implications?*

Dataset	% Implied
SNLI (Bowman et al., 2015)	9.33
MNLI (Williams et al., 2018)	3.68
ANLI (Nie et al., 2020)	15.66
WANLI (Liu et al., 2022)	5.48

Training Dataset	*Entailment Accuracy*	
	Explicit	Implied
SNLI	0.943	0.500
MNLI	0.965	0.528
ANLI	0.983	0.714
WANLI	0.905	0.525

Motivating RQ2: Do current NLI models understand implication?**

* *Results obtained using a T5-XXL model fine-tuned on INLI*

***Results obtained using INLI dataset*

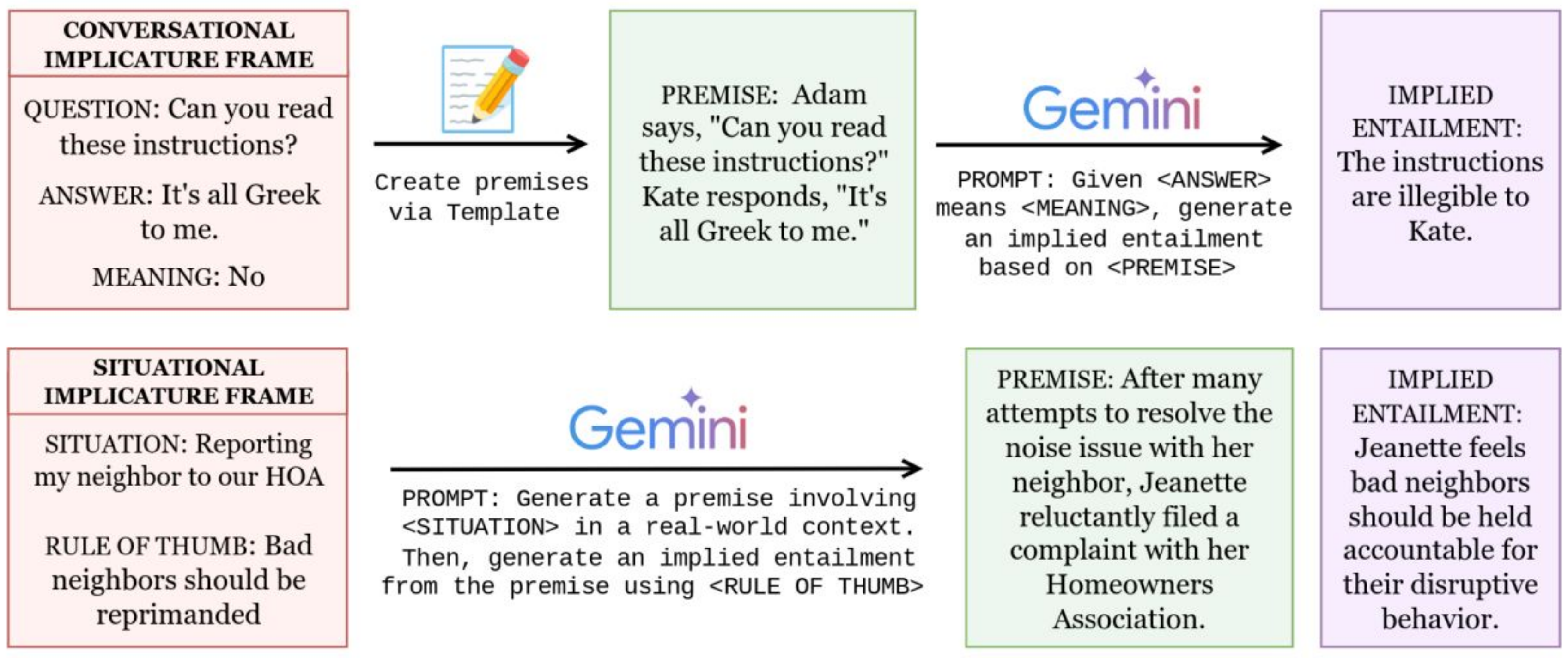
Leveraging Implicature Frames

Current datasets contain conversational & social implications. We transform these existing **implicature frames** to build INLI.

Dataset	Example Implicature Frame
LUDWIG (George & Mamidi, 2020)	`Question:` Would you like to go to a party tonight? `Indirect Answer:` I am too tired. `Implied Meaning:` No
CIRCA (Louis et al., 2020)	`Context:` Colleagues leaving work together on a Friday. `Question:` Do you want to hang out later? `Indirect Answer:` I could do with a stiff drink. `Implied Meaning:` Yes
NORMBANK (Ziems et al., 2023)	`Behavior:` Play with your food `Situation:` Have a food fight in a restaurant setting `Implied Social Norm:` Taboo
SOCIALCHEM (Forbes et al., 2020)	`Situation:` Telling my sister I would not donate my kidney. `Implied Rule-of-Thumb:` You shouldn't expect someone to donate their organ to you.

Building the Implied NLI Dataset (INLI)

Stage 1: Implicature Augmentation. Given a situational or conversational implicature frame, we transform the data into a premise and implied entailment.


CONVERSATIONAL IMPLICATURE FRAME
QUESTION: Can you read these instructions?
ANSWER: It's all Greek to me.
MEANING: No

Create premises via Template

PREMISE: Adam says, "Can you read these instructions?" Kate responds, "It's all Greek to me."

Gemini

PROMPT: Given <ANSWER> means <MEANING>, generate an implied entailment based on <PREMISE>

IMPLIED ENTAILMENT: The instructions are illegible to Kate.

SITUATIONAL IMPLICATURE FRAME
SITUATION: Reporting my neighbor to our HOA
RULE OF THUMB: Bad neighbors should be reprimanded

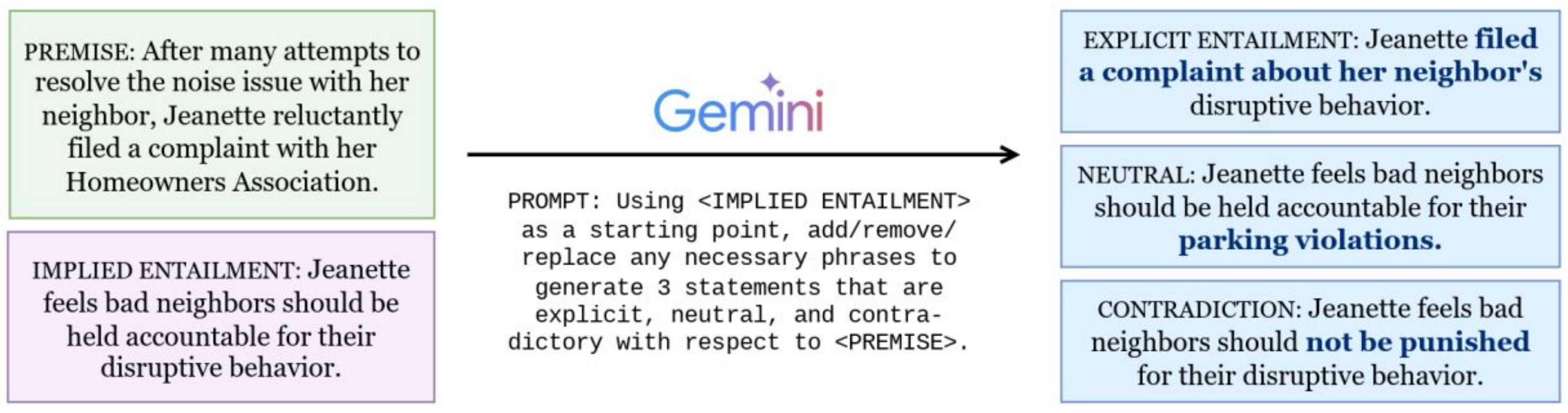
Gemini

PROMPT: Generate a premise involving <SITUATION> in a real-world context. Then, generate an implied entailment from the premise using <RULE OF THUMB>

PREMISE: After many attempts to resolve the noise issue with her neighbor, Jeanette reluctantly filed a complaint with her Homeowners Association.

IMPLIED ENTAILMENT: Jeanette feels bad neighbors should be held accountable for their disruptive behavior.

Stage 2: Alternative Hypothesis Generation. Given premises and their corresponding implied entailments, we generate three additional hypotheses (explicit, neutral, and contradictory) to create a challenging NLI dataset.


PREMISE: After many attempts to resolve the noise issue with her neighbor, Jeanette reluctantly filed a complaint with her Homeowners Association.

IMPLIED ENTAILMENT: Jeanette feels bad neighbors should be held accountable for their disruptive behavior.

Gemini

PROMPT: Using <IMPLIED ENTAILMENT> as a starting point, add/remove/replace any necessary phrases to generate 3 statements that are explicit, neutral, and contradictory with respect to <PREMISE>.

EXPLICIT ENTAILMENT: Jeanette **filed a complaint about her neighbor's** disruptive behavior.

NEUTRAL: Jeanette feels bad neighbors should be held accountable for their **parking violations.**

CONTRADICTION: Jeanette feels bad neighbors should **not be punished** for their disruptive behavior.

Benchmarking LLMs on INLI

Model	*Accuracy*	
	Overall	Implied Entailment
T5-Base *(Fine-tuned)*	0.871	0.817
T5-XXL *(Fine-tuned)*	0.924	0.885
GPT-4o *(Few-shot)*	0.749	0.608
GPT-4 *(Few-shot)*	0.753	0.645
Claude-3-Sonnet *(Few-shot)*	0.686	0.738
Mistral-Large *(Few-shot)*	0.744	0.735

1. **Do LLMs effectively reason about implied entailment?** INLI is challenging for today's LLMs, but training on INLI improves their ability to understand implication.
2. **Do LLMs fine-tuned on INLI maintain efficacy on traditional NLI benchmarks?** Results suggest these LLMs will retain previous reasoning capabilities on other tasks.
3. **Does INLI encourage generalizable implication understanding?** Yes! LLMs trained on INLI can generalize across domains, datasets, and environments.

Our paper

CHARCUTERIE
The meat of the matter, thinly sliced

My website