PizzaCommonSense: Learning to Model Commonsense Reasoning about Intermediate Steps in Cooking Recipes

Aïssatou Diallo, Antonis Bikakis, Luke Dickens, Anthony Hunter, Rob Miller

University College London

EMNLP 2024

Commonsense Reasoning

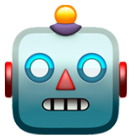
- is a human-like ability to make presumptions about the type and essence of ordinary situations humans encounter every day, incorporating basic facts about actions, their effects, and how the knowledge is obtained.

Place everything in the bowl of an electric mixer with a dough hook.

Ok, what goes in is salt, blue cornmeal, water.

And the output is a nice partially mixed blue dough

The input to this instruction is: "ingredients"

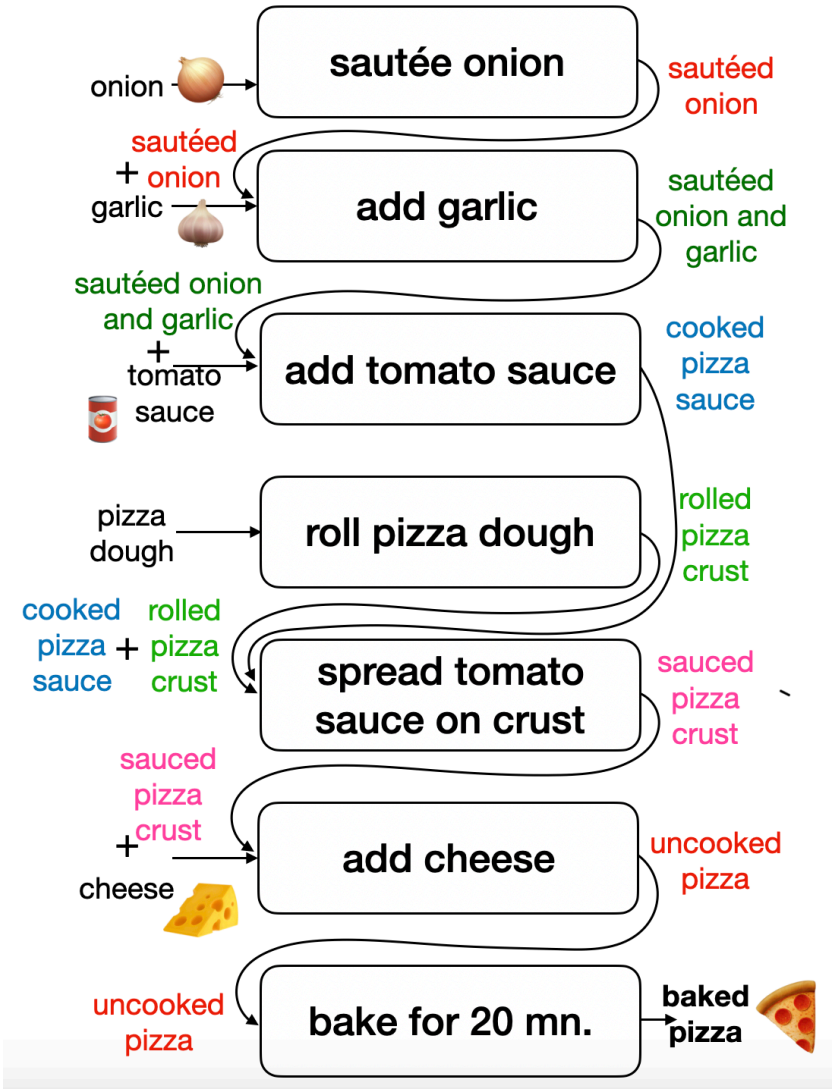


And the output is a "ingredients"

Humans intuitively understand the inputs and outputs of actions, while LLMs may struggle

Objectives

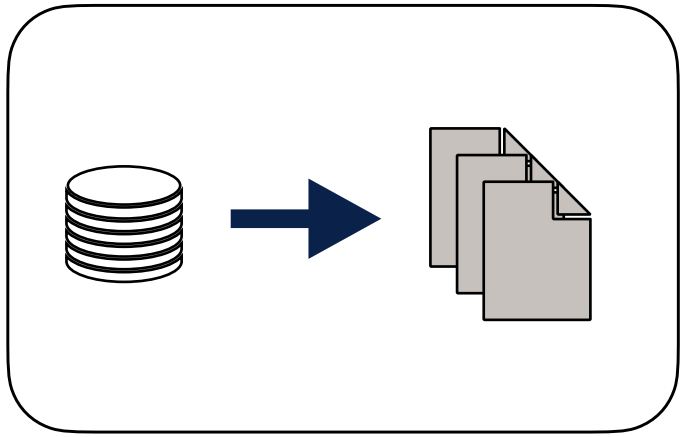
- **PizzaCommonSense**: A dataset capturing intermediate and implicit steps in pizza recipes with detailed input and output comestibles.
- **Enhance Model Reasoning**: Enable models to accurately predict ingredient transformations by resolving implicit references and sequencing actions.
- **Benchmark and Evaluate:** Test models on the dataset to highlight challenges in commonsense reasoning for procedural text.



Why pizza?

- Pizza recipes have balanced complexity and **well-defined steps**, making them ideal for testing procedural comprehension models.
- Pizza preparation involves **distinct and observable ingredient transformations**, aiding the annotation and reasoning of intermediate steps.

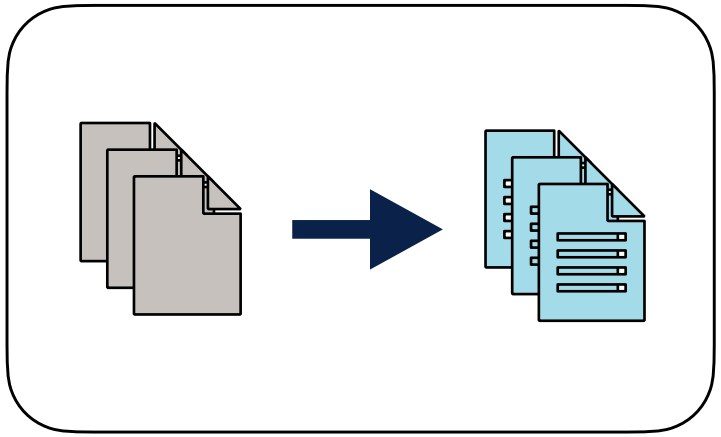
Development Steps



Step 1

Data Collection

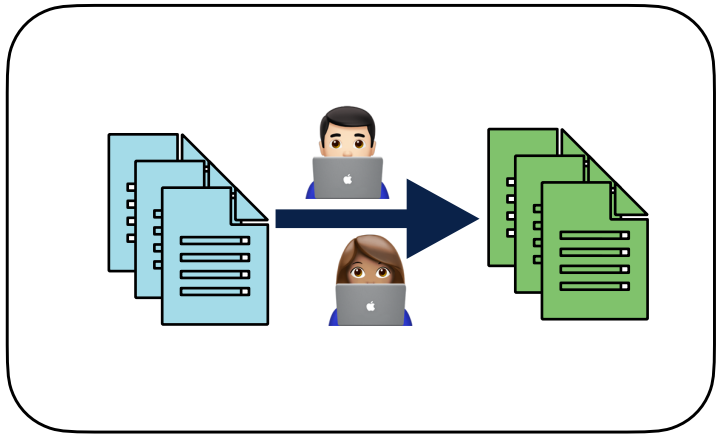
Extract 1,087 unique pizza recipes from the Recipe1M dataset.



Step 2

Pre-processing

Split recipes into atomic instructions and identify main actions using a glossary.

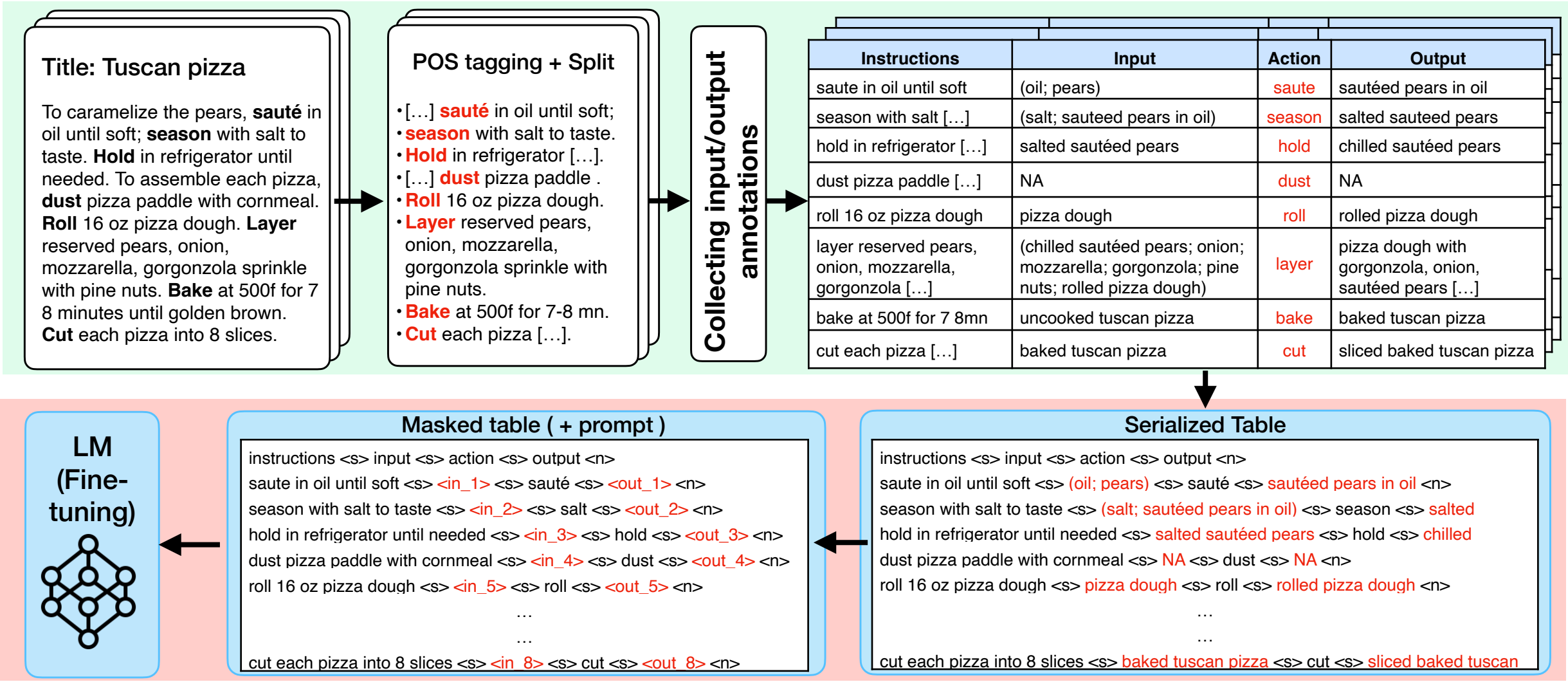


Step 3

Annotation

AMT workers label each instruction's "Input" and "Output" comestibles.

PizzaCommonSense



Experimental Baselines

Models:

- T5
- Flan-T5
- GPT3.5 + 1-shot
- GPT3.5 + CoT
- GPT3.5 + Fine Tuning
- GPT4 + CoT
- + Human

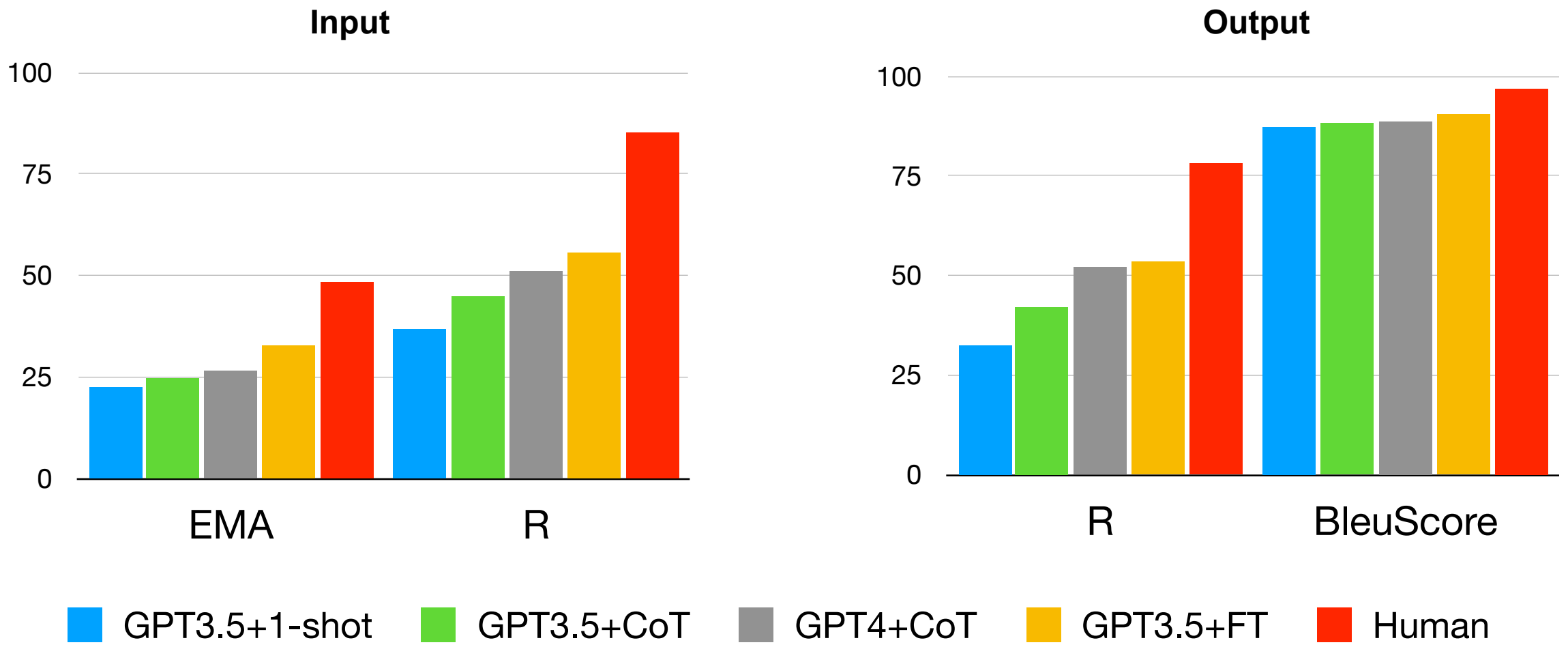
Metrics:

- Rouge
- Exact Matching Accuracy (EMA)
- BleuScore

Human Evaluation

- Completeness: true if all the relevant comestibles are present in the generated input;
- Validity: if the generated text follows all the rules ("only comestibles", "descriptive predictions”);
- Consistency: the input-output pair makes logical sense;
- Win/Tie/Lose: the generated outputs are compared to the gold reference to determine whether they are preferred (Win), equivalent (Tie), or less preferred (Lose).

Experimental Baselines



Experimental Baselines

	Completeness	Validity	Consistency	Win (↑)	Tie	Lose (↓)
Flan T5 + FT	0.52	0.42	0.82	0.04	0.04	0.92
GPT-3.5 + FT	**0.72**	**0.72**	**1.0**	0.24	0.02	0.74
GPT-4 + CoT	0.68	0.68	0.95	0.26	0.06	0.68
Human	0.95	0.98	1.0	0.3	0.65	0.05

Analysis

Instructions		Ground-truth	T5	Flan T5	GPT3.5	GPT4
combine all ingredients except for the chicken, oil and cheese in a saucepan.	I	(tomato sauce; water; oregano; ...)	(~~chicken~~; ~~oil~~; ~~cheese~~)	(all ingredient; ~~oil~~; ~~cheese~~)	tomatoes, onion, garlic, oregano,...	(tomato; salt; italian seasoning)
	O	seasoned tomato sauce	sauce	~~chicken fat~~	combined tomato sauce mixture	tomato sauce
add more flour or water until you get that consistency	I	(opt. flour; opt. water; dough)	(dough; flour; water)	(~~mixed ingredients~~; flour; water)	flour or water	(dough; water or flour)
	O	dough	flour added to dough	flour or water added to mixed ingredients	~~adjusted consistency~~	water or flour added to dough
place everything in the bowl of an electric mixer with a dough hook	I	(...; salt; blue cornmeal)	(dough; olive oil)	(~~peanut butter; fresh dill~~; ...)	ingredients	ingredients
	O	partially mixed blue dough	dough	~~peanut butter mixture~~	ingredients in the bowl	ingredients
before putting the chicken on, get a fork	I	NA	(chicken; fork)	NA	NA	(~~fork~~; chicken)
	O	NA	chicken with ~~fork~~	NA	NA	chicken and ~~fork~~

Conclusion

- To show **Complete Process Understanding and Ingredient Transformation**
- Identify intermediates to help models infer **implicit information** for better comprehension.
- Understand intermediates to accurately predict the order and dependencies of cooking actions for **better sequential reasoning.**

Thank you

Paper : arxiv.org/abs/2401.06930

Code/Data : github.com/adiallo07/PizzaCommonsense