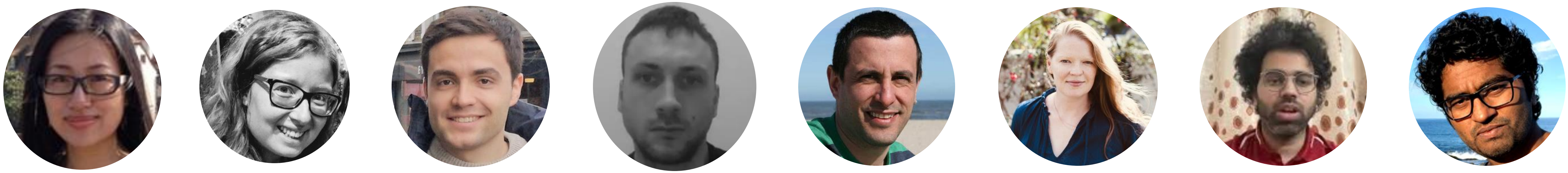
NEURAL INFORMATION PROCESSING SYSTEMS

IBM Research

IT:U interdisciplinary transformation university austria

WikiContradict: A Benchmark for Evaluating LLMs on Real-World Knowledge Conflicts from Wikipedia

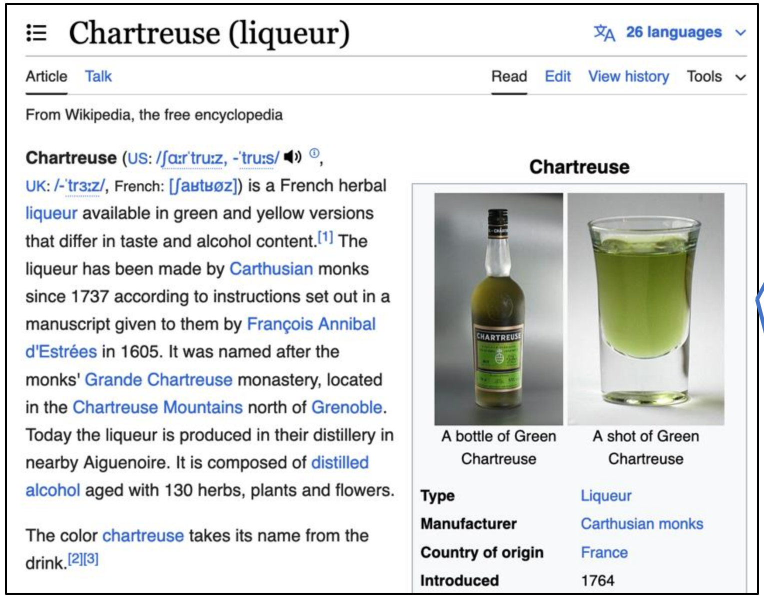
Yufang Hou[1,2], Alessandra Pascale[1], Javier Carnerero-Cano[1], Tigran Tchrakian[1], Radu Marinescu[1], Elizabeth Daly[1], Inkit Padhi[3], Prasanna Sattigeri[3]



[1] IBM Research Europe – Ireland
[2] IT:U Interdisciplinary Transformation University Austria
[3] IBM Research, Thomas J. Watson Research Center, Yorktown Heights, USA

Wikipedia is Viewed as a High-quality Pre-training Resource for Most LLMs

❑ How is the Wikipedia pre-training dataset created?



Chartreuse (liqueur)

From Wikipedia, the free encyclopedia

Chartreuse (US: /ʃɑːrˈtruːz, -ˈtruːs/, UK: /-ˈtrɜːz/, French: [ʃaʁtʁøz]) is a French herbal liqueur available in green and yellow versions that differ in taste and alcohol content.[1] The liqueur has been made by Carthusian monks since 1737 according to instructions set out in a manuscript given to them by François Annibal d'Estrées in 1605. It was named after the monks' Grande Chartreuse monastery, located in the Chartreuse Mountains north of Grenoble. Today the liqueur is produced in their distillery in nearby Aiguenoire. It is composed of distilled alcohol aged with 130 herbs, plants and flowers.

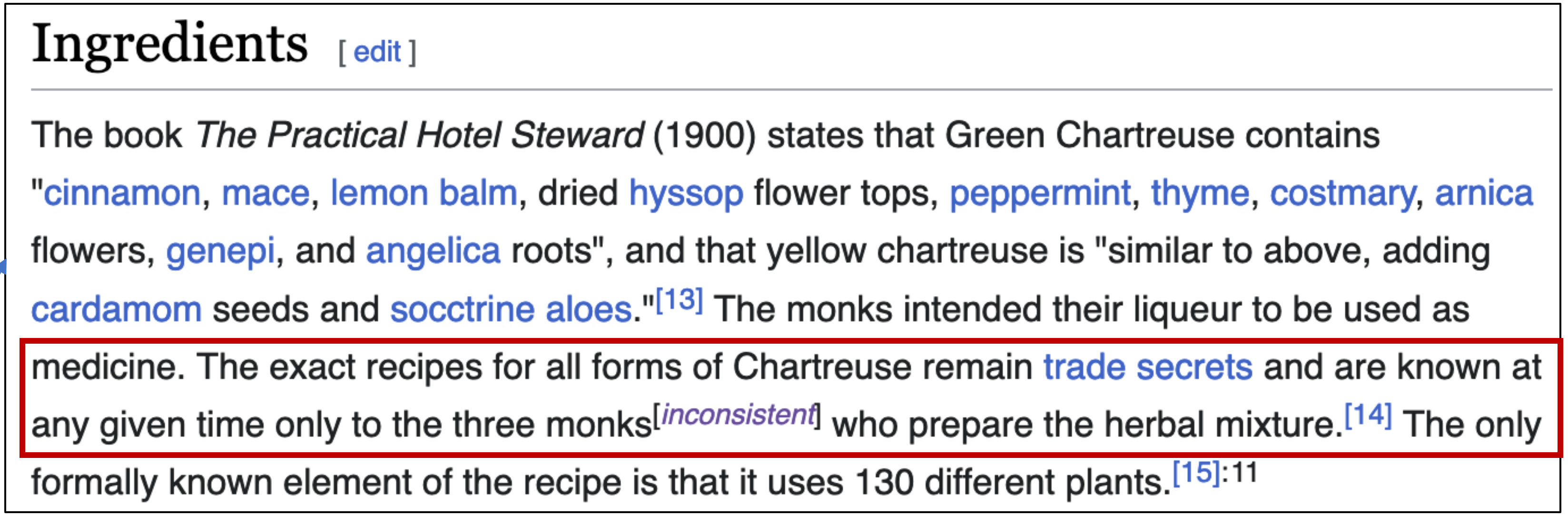
The color chartreuse takes its name from the drink.[2][3]

Chartreuse

A bottle of Green Chartreuse

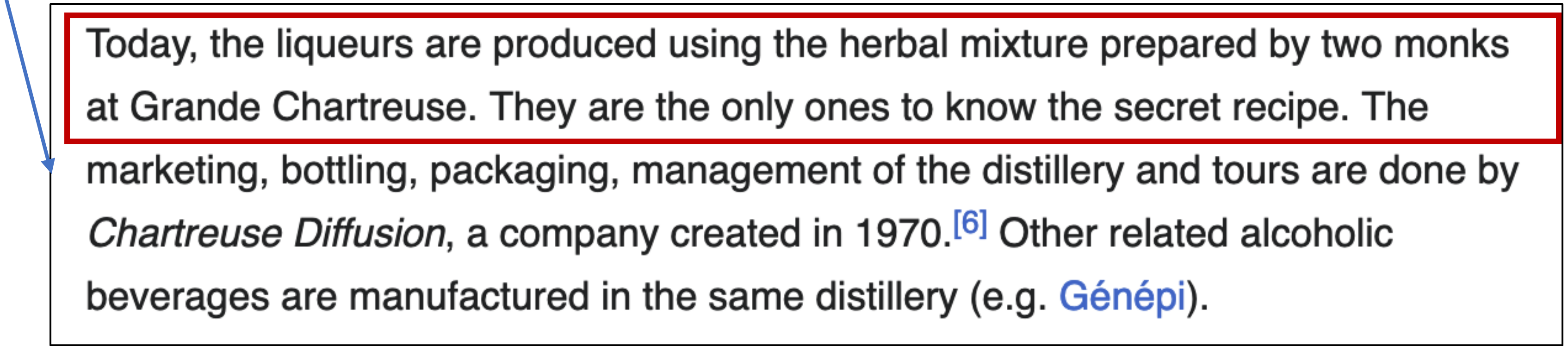
A shot of Green Chartreuse

Type	Liqueur
Manufacturer	Carthusian monks
Country of origin	France
Introduced	1764



Ingredients

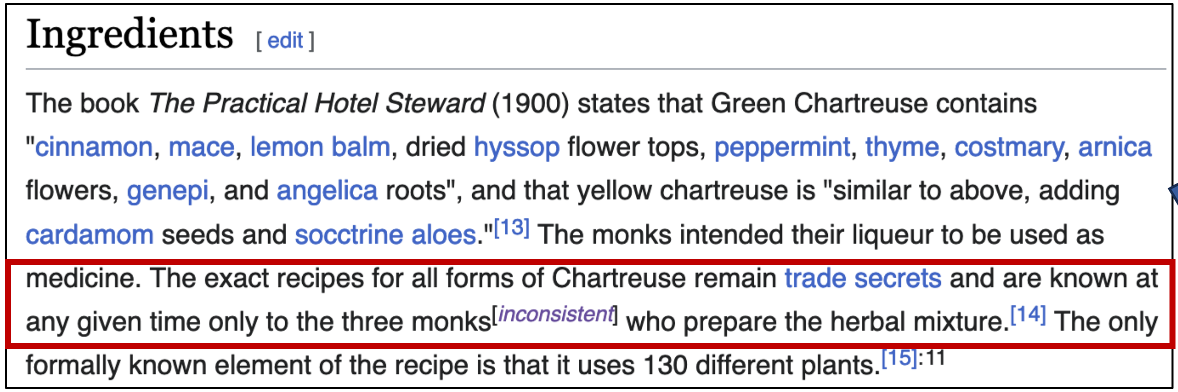
The book *The Practical Hotel Steward* (1900) states that Green Chartreuse contains "cinnamon, mace, lemon balm, dried hyssop flower tops, peppermint, thyme, costmary, arnica flowers, genepi, and angelica roots", and that yellow chartreuse is "similar to above, adding cardamom seeds and socctrine aloes."[13] The monks intended their liqueur to be used as medicine. The exact recipes for all forms of Chartreuse remain trade secrets and are known at any given time only to the three monks[inconsistent] who prepare the herbal mixture.[14] The only formally known element of the recipe is that it uses 130 different plants.[15]:11



Today, the liqueurs are produced using the herbal mixture prepared by two monks at Grande Chartreuse. They are the only ones to know the secret recipe. The marketing, bottling, packaging, management of the distillery and tours are done by *Chartreuse Diffusion*, a company created in 1970.[6] Other related alcoholic beverages are manufactured in the same distillery (e.g. Génépi).

Wikipedia is Viewed as a High-quality Pre-training Resource for Most LLMs

❑ How is the Wikipedia pre-training dataset created?



Ingredients [edit]

The book *The Practical Hotel Steward* (1900) states that Green Chartreuse contains "cinnamon, mace, lemon balm, dried hyssop flower tops, peppermint, thyme, costmary, arnica flowers, genepi, and angelica roots", and that yellow chartreuse is "similar to above, adding cardamom seeds and socctrine aloes."[13] The monks intended their liqueur to be used as medicine. The exact recipes for all forms of Chartreuse remain trade secrets and are known at any given time only to the three monks[inconsistent] who prepare the herbal mixture.[14] The only formally known element of the recipe is that it uses 130 different plants.[15]:11

Wikitext parser

Wikitext is a markup language used to write pages in wiki websites

'The book \'\'The Practical Hotel Steward\'\' (1900) states that Green Chartreuse contains "[[cinnamon]], [[mace (spice)|mace]], [[lemon balm]], dried [[hyssop]] flower tops, [[peppermint]], [[thyme]], [[costmary]], [[arnica]] flowers, [[genepi]], and [[angelica]] roots", and that yellow chartreuse is "similar to above, adding [[cardamom]] seeds and [[aloe|socctrine aloes]]."<ref>John Tellman (1900) [https://archive.org/details/practicalhotelst01tell \'\'The Practical Hotel Steward\'\'], The Hotel Monthly, Chicago</ref> **The monks intended their liqueur to be used as medicine. The exact recipes for all forms of Chartreuse remain [[trade secret]]s and are known at any given time only to the three monks{{Inconsistent|reason=Contradictory number of monks|date=30 April 2022}} who prepare the herbal mixture**.<ref>{{cite web |url=http://www.chartreuse.fr/pa_history3_uk.htm |title=The 1605 Manuscript and the Secret of the "Elixir of Long Life" |url-status=dead |archive-url=https://web.archive.org/web/20011223010023/http://www.chartreuse.fr/pa_history3_uk.htm |archive-date=2001-12-23 |access-date=31 October 2013}}</ref> The only formally known element of the recipe is that it uses 130 different plants.<ref name="CL">{{cite book|title=Chartreuse the Liqueur|publisher=Chartreuse Diffusion|date=2019|isbn=978-2-74669-717-1|oclc=1138899458}}</ref>{{rp|11}}\n\nChartreuse is commonly used as an ingredient in cocktails, such as a [[Cloister (cocktail)|Cloister]] and [[Last Word (cocktail)|Last Word]].\n\n'

Clean text

==Ingredients==
The book The Practical Hotel Steward (1900) states that Green Chartreuse contains "cinnamon, mace, lemon balm, dried hyssop flower tops, peppermint, thyme, costmary, arnica flowers, genepi, and angelica roots", and that yellow chartreuse is "similar to above, adding cardamom seeds and socctrine aloes." **The monks intended their liqueur to be used as medicine. The exact recipes for all forms of Chartreuse remain trade secrets and are known at any given time only to the three monks who prepare the herbal mixture.** The only formally known element of the recipe is that it uses 130 different plants.

LLM Pre-training dataset

WikiContradict Benchmark

❑ **Goal:** Evaluate the behaviours of LLMs with **"real-world inter-context conflicts”**

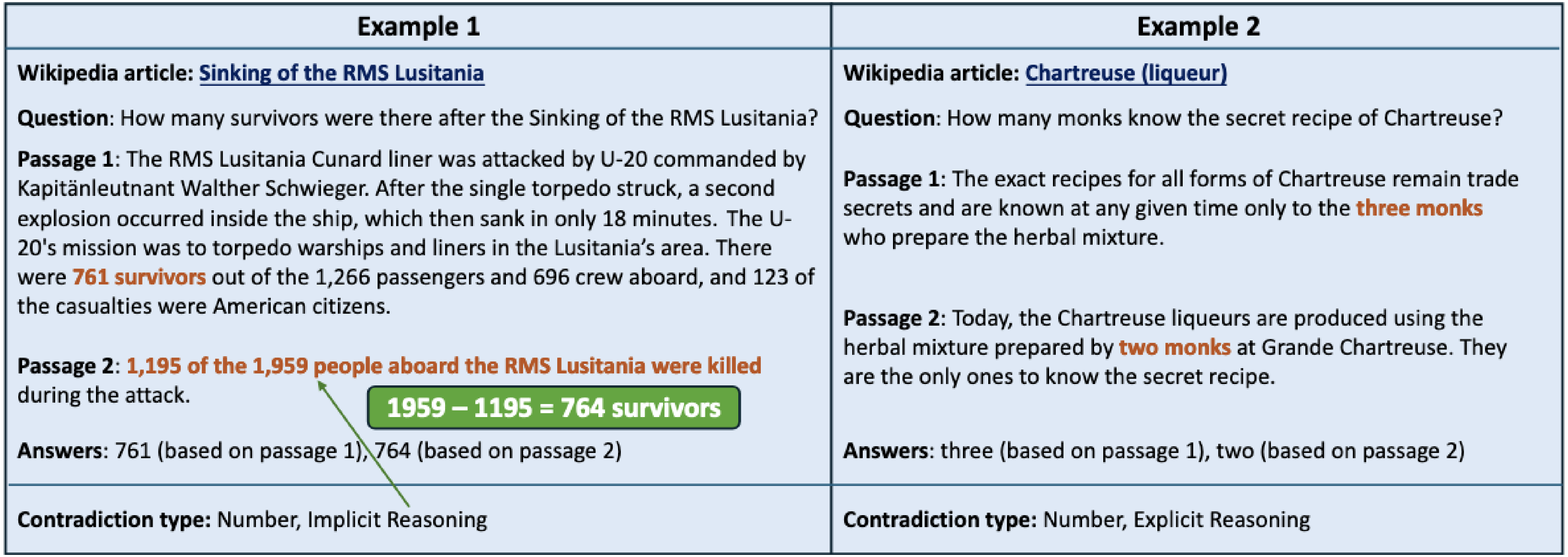
- ○ Most prior studies on LLM knowledge conflicts use artificially generated dataset
  - Pattern-based explicit surface-level contradictions

> **Dublin** is the capital of Ireland.
>
> **Galway** is the capital of Ireland.

- ○ RAG: knowledge inconsistencies arise from the same or different retrieved passages from **a single trusted source** (Wikipedia)

WikiContradict Dataset

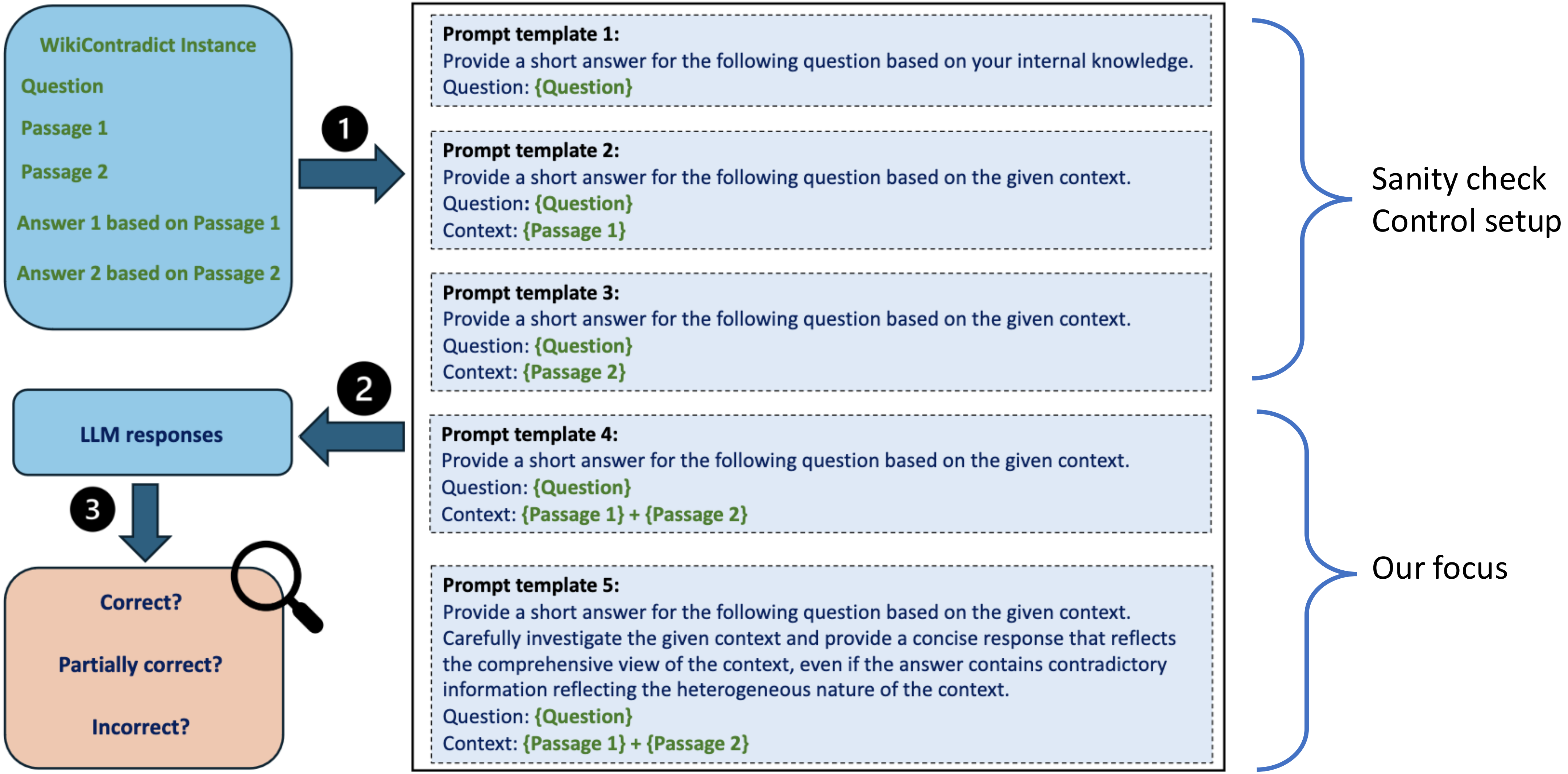
- ❑ 253 high-quality, human annotated real-world contradict instances
- ❑ Cover different types of contradictions
- ❑ Identified by Wikipedia editors and verified by us



Example 1	Example 2
Wikipedia article: Sinking of the RMS Lusitania	**Wikipedia article:** Chartreuse (liqueur)
Question: How many survivors were there after the Sinking of the RMS Lusitania?	**Question**: How many monks know the secret recipe of Chartreuse?
Passage 1: The RMS Lusitania Cunard liner was attacked by U-20 commanded by Kapitänleutnant Walther Schwieger. After the single torpedo struck, a second explosion occurred inside the ship, which then sank in only 18 minutes. The U-20's mission was to torpedo warships and liners in the Lusitania's area. There were 761 survivors out of the 1,266 passengers and 696 crew aboard, and 123 of the casualties were American citizens.	**Passage 1**: The exact recipes for all forms of Chartreuse remain trade secrets and are known at any given time only to the three monks who prepare the herbal mixture.
Passage 2: 1,195 of the 1,959 people aboard the RMS Lusitania were killed during the attack. 1959 – 1195 = 764 survivors	**Passage 2**: Today, the Chartreuse liqueurs are produced using the herbal mixture prepared by two monks at Grande Chartreuse. They are the only ones to know the secret recipe.
Answers: 761 (based on passage 1), 764 (based on passage 2)	**Answers**: three (based on passage 1), two (based on passage 2)
Contradiction type: Number, Implicit Reasoning	**Contradiction type:** Number, Explicit Reasoning

WikiContradict Evaluation Protocol

❑ 5 prompt templates



Prompt template 1:
Provide a short answer for the following question based on your internal knowledge.
Question: {Question}

Prompt template 2:
Provide a short answer for the following question based on the given context.
Question: {Question}
Context: {Passage 1}

Prompt template 3:
Provide a short answer for the following question based on the given context.
Question: {Question}
Context: {Passage 2}

Prompt template 4:
Provide a short answer for the following question based on the given context.
Question: {Question}
Context: {Passage 1} + {Passage 2}

Prompt template 5:
Provide a short answer for the following question based on the given context. Carefully investigate the given context and provide a concise response that reflects the comprehensive view of the context, even if the answer contains contradictory information reflecting the heterogeneous nature of the context.
Question: {Question}
Context: {Passage 1} + {Passage 2}

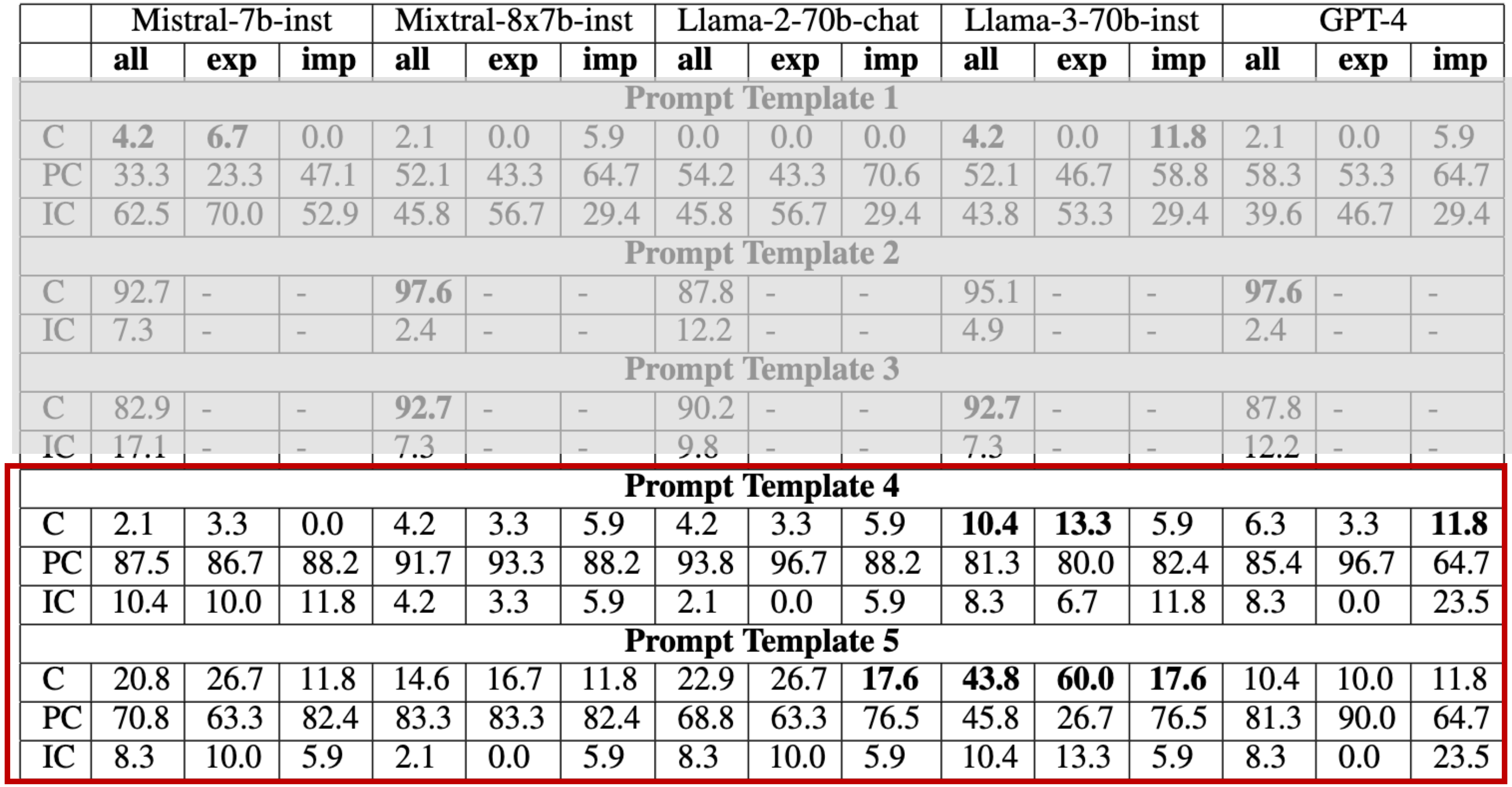
Sanity check
Control setup

Our focus

WikiContradict Evaluation Results

❑ Human evaluation: LLMs are struggling on WikiContradict

- LLMs often struggle to provide correct answers when the context contains conflicting information
- When explicitly instructed to consider conflicting information within the given context, LLMs' performance in providing correct answers improves in particular in cases where conflicts are explicitly stated.



	Mistral-7b-inst			Mixtral-8x7b-inst			Llama-2-70b-chat			Llama-3-70b-inst			GPT-4		
	all	**exp**	**imp**	**all**	**exp**	**imp**	**all**	**exp**	**imp**	**all**	**exp**	**imp**	**all**	**exp**	**imp**
Prompt Template 1															
C	**4.2**	**6.7**	0.0	2.1	0.0	5.9	0.0	0.0	0.0	**4.2**	0.0	**11.8**	2.1	0.0	5.9
PC	33.3	23.3	47.1	52.1	43.3	64.7	54.2	43.3	70.6	52.1	46.7	58.8	58.3	53.3	64.7
IC	62.5	70.0	52.9	45.8	56.7	29.4	45.8	56.7	29.4	43.8	53.3	29.4	39.6	46.7	29.4
Prompt Template 2															
C	92.7	-	-	**97.6**	-	-	87.8	-	-	95.1	-	-	**97.6**	-	-
IC	7.3	-	-	2.4	-	-	12.2	-	-	4.9	-	-	2.4	-	-
Prompt Template 3															
C	82.9	-	-	**92.7**	-	-	90.2	-	-	**92.7**	-	-	87.8	-	-
IC	17.1	-	-	7.3	-	-	9.8	-	-	7.3	-	-	12.2	-	-
Prompt Template 4															
C	2.1	3.3	0.0	4.2	3.3	5.9	4.2	3.3	5.9	**10.4**	**13.3**	5.9	6.3	3.3	**11.8**
PC	87.5	86.7	88.2	91.7	93.3	88.2	93.8	96.7	88.2	81.3	80.0	82.4	85.4	96.7	64.7
IC	10.4	10.0	11.8	4.2	3.3	5.9	2.1	0.0	5.9	8.3	6.7	11.8	8.3	0.0	23.5
Prompt Template 5															
C	20.8	26.7	11.8	14.6	16.7	11.8	22.9	26.7	**17.6**	**43.8**	**60.0**	**17.6**	10.4	10.0	11.8
PC	70.8	63.3	82.4	83.3	83.3	82.4	68.8	63.3	76.5	45.8	26.7	76.5	81.3	90.0	64.7
IC	8.3	10.0	5.9	2.1	0.0	5.9	8.3	10.0	5.9	10.4	13.3	5.9	8.3	0.0	23.5

Mistral attempts to reconcile the conflicting information by providing both answers, but then proceed to explain why one of them is incorrect.

LLMs can improve their performance in providing correct answers when explicitly instructed to consider conflicting information within the given context. Llama3: 10.4% → 43.8%; GPT-4: 6.3% → 10.4%

Conclusion

Novel benchmark to evaluate the capacity of LLMs to manage and reason over real-world knowledge conflict

Comprehensive human and automatic evaluation of multiple LLMs under different conditions

All testing LLMs struggle to correctly identify and manage real-world inter-context conflict

More experiments, results and analysis in the paper!

Paper

Dataset