

What Has Been Lost with Synthetic Evaluation?

Alexander Gill et al., 2025 | University of Utah &

Background & Motivation

Benchmarks are **the backbone of NLP progress** — they define what we measure.

Background & Motivation

Benchmarks are **the backbone of NLP progress** — they define what we measure.

A good benchmark should:

- Target **specific** reasoning capabilities
- Avoid exploitable **shortcuts**
- Be **challenging** even for frontier LLMs

Background & Motivation

Benchmarks are **the backbone of NLP progress** — they define what we measure.

A good benchmark should:

- Target **specific** reasoning capabilities
- Avoid exploitable **shortcuts**
- Be **challenging** even for frontier LLMs

Crowdsourcing is expensive and inconsistent

while LLMs are **cheap, fast, and controllable**

→ tempting alternative 

Research Question

When created under the same annotation criteria,

Do LLM-generated benchmarks match human-authored ones?

- **Validity**
- **Difficulty**
- **Model ranking**

Case Studies

Two reasoning-over-text benchmarks:

Dataset	Focus	Example Reasoning Type
CondaQA	Reasoning about negation	Scope / Paraphrase / Affirmative edits
DROP	Numerical & compositional reasoning	Compositional / Temporal edits

Case Studies

CondaQA: Reasoning about Negation

Passage: "The package was delivered on Monday, but not on Tuesday."

Case Studies

CondaQA: Reasoning about Negation

Passage: "The package was delivered on Monday, but not on Tuesday."

A simple, "bad" question: "Was the package delivered on Tuesday?"

Case Studies

CondaQA: Reasoning about Negation

Passage: "The package was delivered on Monday, but not on Tuesday."

A simple, "bad" question: "Was the package delivered on Tuesday?"

A CondaQA-style, "good" question:

"If you waited at home for the package on Tuesday, would you have received it?

Case Studies

DROP: Reasoning over Quantities

Passage: "Team A scored 3 goals in the first half and 2 goals in the second half. Team B scored a total of 4 goals in the game."

Case Studies

DROP: Reasoning over Quantities

Passage: "Team A scored 3 goals in the first half and 2 goals in the second half. Team B scored a total of 4 goals in the game."

A simple, "bad" question: "How many goals did Team B score?"

Case Studies

DROP: Reasoning over Quantities

Passage: "Team A scored 3 goals in the first half and 2 goals in the second half. Team B scored a total of 4 goals in the game."

A simple, "bad" question: "How many goals did Team B score?"

A DROP-style, "good" question:

"How many more goals did Team A score than Team B?"

Prior Case Studies

CondaQA	Human [†]	o3-mini [△]	Llama-3.3°
Original-Data Accuracy	91.9	72.4	78.8
Full Bundle Consistency	81.6	45.9	48.5
Paraphrase Consistency	93.6	69.6	76.5
Scope Consistency	86.5	55.5	62.8
Affirmative Consistency	88.2	56.9	67.3

DROP	Human [‡]	o3-mini [△]	Llama-3.3°
Original-Data Token F1	96.4	84.3	70.9
Consistency	N/A	53.5	35.5

Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 2022. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation.

These questions remain hard even for o3-mini or Llama-3.3
→ ideal for testing synthetic generation

Method: Synthetic Data Generation Pipeline

Input: original Wikipedia passage

Prompt **GPT-4-Turbo (2024-04-09)** to generate:

- new questions
- contrastive edits (paraphrase, scope, affirmative, etc.)

Iterative prompt refinement + filtering until $\geq 85\%$ validity

Manual validation of hundreds of samples

→ Overall **421 passage-question pairs** and **926 edits** tested

Prompting Strategies

Prompting Strategy	Description	Data
Overgeneration	Generating more examples than will be retained in the final dataset	CondaQA Questions, CondaQA Edits, DROP Questions, DROP Edits
Filters and Verifiers	Using LLMs or other automated methods to assess generated examples for specific quality criteria, and filter out examples that do not meet those criteria	CondaQA Questions, CondaQA Edits, DROP Questions, DROP Edits
Reasoning Step Types	Prompt includes examples that capture the reasoning type in question, model is asked to generate an example that encapsulates that type	DROP Questions, DROP Edits
Human Guidelines	Use annotation instructions given to human annotators	CondaQA Questions, CondaQA Edits, DROP Questions, DROP Edits
Decomposing	Breaking down the annotation task into sub-tasks an LLM can follow	CondaQA Questions, CondaQA Edits, DROP Questions, DROP Edits
QA Generation	Help an LLM understand whether a piece of information changed by having it generate questions for which answers should change if the change was made	CondaQA scope edit

Table 10: An overview of prompting strategies used in the study (Prompting Strategy), descriptions of the prompting ideas (Description), and the datasets they were used in generating (Data).

Prompt Engineering Insights

- Human annotation guidelines ≠ good prompts
 - Effective prompting strategies:
 - Decompose task into sub-steps
 - Over-generate & filter improves quality
 - **But:** strategies that work for one dataset often fail for another
- Prompt validity ≈ high cost in manual iteration

Validity Results from one of Author

Dataset / Task	Validity (%)
CondaQA Questions	72.8
CondaQA Edits (Paraphrase / Scope / Affirmative)	96.5 / 88.6 / 98.2
DROP Questions	88.7
DROP Edits (Compositional / Temporal)	65 / 86

Validity Criteria

Annotation Item	Validity Criteria
CONDAQA	
<i>Question</i>	The question (i) targets the negated statement, rather than other information in the provided passage, and (ii) is about an implication of the negated statement.
<i>Paraphrase Edit</i>	The edit (i) yields a passage with the same meaning as the original passage, (ii) is coherent, and (iii) does not include the original negation cue.
<i>Scope Edit</i>	The edit (i) changes what is being negated by the negation cue, and (ii) is coherent.
<i>Affirmative Edit</i>	The edit (1) changes the passage such that what was being negated in the original passage is no longer negated, and (2) is coherent.
DROP	
<i>Question</i>	The question (i) requires discrete reasoning over the passage and (ii) is not directly answerable from the passage.
<i>Compositional Edit</i>	The edit adds additional reasoning step(s) to the original question.
<i>Temporal Edit</i>	The edit modifies either the order in which events appear in the original passage or the dates associated with each event to swap their temporal order.

Table 11: Validity criteria for each annotation item.

Validity Results from External Experts



User study: 16 NLP researchers (PhDs)

Which one better matched the dataset specification?

Validity Results

Synthetic samples are
preferred in general

	CondaQA	% Valid (Annotators)	% Valid (Author)
<i>Gen.</i>	Paraphrase	77.4	90.3
	Scope	64.5	87.1
	Affirmative	83.9	93.5
	Questions	64.5	77.4
<i>Orig.</i>	Paraphrase	64.5	-
	Scope	61.3	-
	Affirmative	90.3	-
	Questions	58.1	-
DROP			
<i>Gen.</i>	Compositional	78.1	53.1
	Questions	81.2	84.4
<i>Orig.</i>	Compositional	84.4	-
	Questions	78.1	-

But Validity ≠ Difficulty

But Validity ≠ Difficulty

When benchmarked on CondaQA & DROP

Using **6 LLM families** (GPT-4, Claude, Gemini, Llama, Qwen etc.)

Synthetic data is easier and less discriminative.

But Validity ≠ Difficulty

- Scores on synthetic benchmarks ↑ significantly across most models
- Human datasets still harder and rank models differently

	GPT-4-Turbo [△]		GPT-4o [°]		o3-mini [†]		Claude-Opus-4 [°]		Gemini-2.5-Flash [○]		Llama-3.3-70b [‡]		Qwen2.5-72b [*]	
	Orig.	Gen.	Orig.	Gen.	Orig.	Gen.	Orig.	Gen.	Orig.	Gen.	Orig.	Gen.	Orig.	Gen.
CondaQA														
Original-Data Accuracy	73.9 _{2.6}	79.7_{1.2}	70.5 _{0.9}	78.3_{1.4}	68.5 _{1.9}	83.4_{1.4}	67.5 _{2.5}	80.3_{0.9}	70.2 _{0.9}	81.7_{1.4}	81.4 _{0.0}	81.4 _{0.0}	69.5 _{0.0}	78.3_{0.8}
Full Bundle Consistency	40.5 _{2.7}	44.5_{4.1}	32.6 _{1.6}	50.9_{5.2}	40.5 _{2.1}	47.3_{3.0}	41.9 _{2.8}	44.1_{5.0}	54.0_{3.4}	39.5 _{2.6}	53.0_{3.0}	40.9 _{2.3}	40.9 _{2.1}	44.5_{4.7}
Paraphrase Consistency	69.5 _{1.5}	73.9_{1.0}	66.5 _{1.6}	72.5_{1.0}	65.5 _{3.1}	78.6_{1.8}	64.7 _{1.6}	72.1_{3.0}	69.1 _{1.8}	72.5_{2.7}	78.9_{1.6}	77.1 _{1.5}	70.2 _{1.0}	74.6_{1.5}
Scope Consistency	53.3 _{2.4}	67.8_{1.7}	44.2 _{0.9}	64.5_{3.1}	49.2 _{5.6}	65.3_{2.0}	49.2 _{3.2}	62.0_{4.0}	59.6 _{2.4}	64.5_{1.1}	60.0_{2.7}	58.8 _{1.7}	49.2 _{2.4}	59.6_{3.4}
Affirmative Consistency	55.9 _{4.0}	56.2_{2.1}	46.9 _{3.2}	63.8_{1.6}	51.8 _{2.3}	61.9_{3.4}	47.8 _{2.7}	56.2_{3.4}	59.6_{3.4}	53.1 _{3.2}	68.6_{1.8}	60.0 _{0.9}	54.7 _{1.7}	58.1_{2.1}
DROP														
Original-Data Token F1	61.7 _{2.1}	83.0_{3.1}	62.8 _{2.6}	83.9_{1.7}	75.2 _{4.1}	92.7_{1.3}	73.2 _{0.8}	90.3_{1.4}	71.9 _{1.8}	90.9_{0.0}	52.8 _{2.0}	75.1_{2.8}	57.1 _{1.8}	77.9_{2.0}
Compositional Consistency	30.4 _{3.6}	58.8_{2.7}	29.2 _{3.0}	63.6_{0.9}	57.2 _{3.0}	80.8_{3.0}	55.6 _{2.6}	82.0_{2.4}	58.0 _{2.4}	86.4_{1.7}	25.2 _{1.1}	35.2_{3.3}	23.2 _{2.3}	48.4_{3.6}

Key Findings

Dimension	Synthetic Benchmark	Human Benchmark
Validity	High (formally correct)	High with nuance
Difficulty	Low (easy for LLMs)	High (natural complexity)
Model Ranking	Unstable / distorted	Reliable
Perceived Quality	"Looks good"	Less consistent but richer
Cost & Speed	Very low & fast	Expensive & slow

What Has Been Lost

“What has been lost” = challenge >> nuance and creativity

LLMs follow rules but lack human ingenuity.

Synthetic datasets under-estimate model limitations.

**Human annotation remains essential
for probing deep reasoning and real-world generalization.**

Practical Implications

- Use synthetic evaluations for  **unit tests & regression checks.**
- **Retain human evaluation** for  benchmarking intelligence and robustness.
- Future direction: combine human feedback with LLM generation loops for
“difficulty-aware” benchmark construction

Takeaways

LLMs can generate valid but not challenging evaluation data.

Synthetic benchmarks may inflate model performance and distort ranking.

Human creativity remains indispensable for measuring reasoning capability.



**Utrecht
University**

Sharing science,
shaping tomorrow