MATHWELL: Generating Educational Math Word Problems at Scale

Bryan Christ, Jonathan Kropko, Tom Hartvigsen

Contact: brc4cb@virginia.edu

UVA SCHOOL *of* DATA SCIENCE

TLDR; We introduce context-free educational grade school math word problem generation and release a large dataset for this task.

Word Problems are Critical K-8 Educational Tools

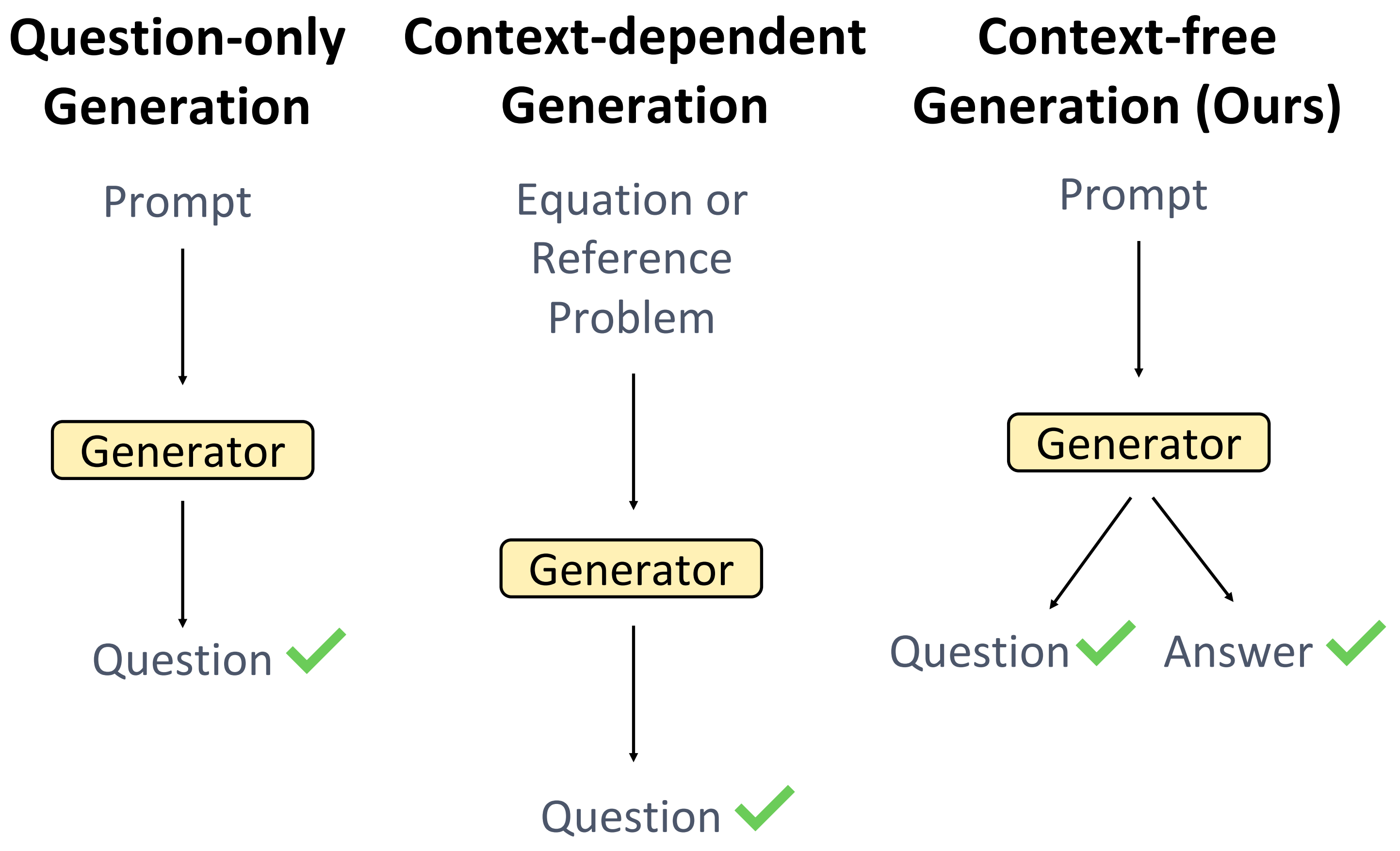
Math word problems assess the highest level of student knowledge and customizing them promotes student learning. However, writing these problems:

Is time consuming

Requires domain expertise

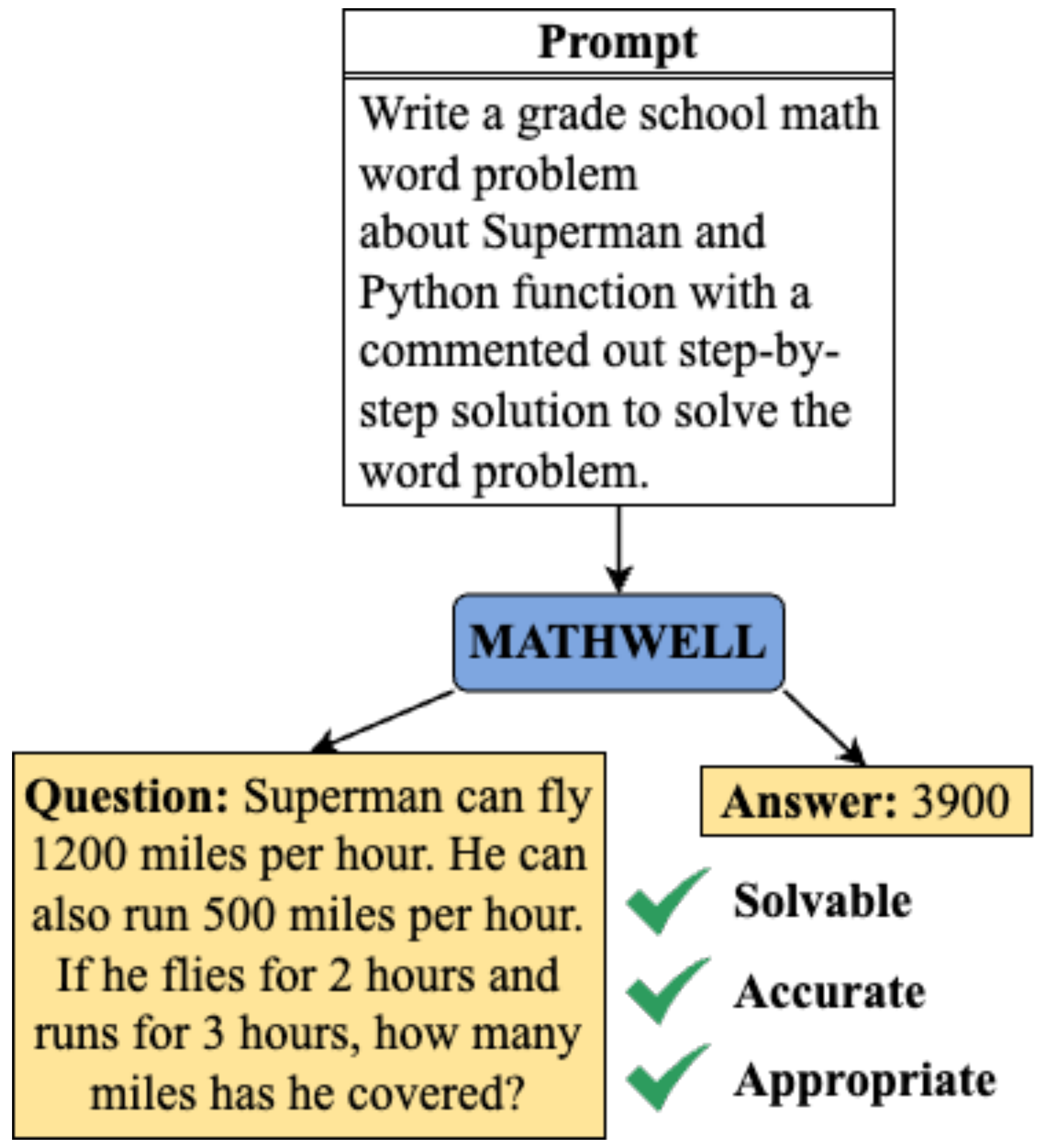
We need methods to automatically generate customized math word problems

Comparison with Prior Works



Existing methods require too much manual curation to be useful for teachers. Therefore, we propose *context-free* educational math word problem generation.

First Context-free Word Problem Generator: MATHWELL



Prompt

Write a grade school math word problem about Superman and Python function with a commented out step-by-step solution to solve the word problem.

MATHWELL

Question: Superman can fly 1200 miles per hour. He can also run 500 miles per hour. If he flies for 2 hours and runs for 3 hours, how many miles has he covered?

Answer: 3900

✓ **Solvable**
✓ **Accurate**
✓ **Appropriate**

MATHWELL generates customized educational math word problems and Python function solutions to these problems without the need for a pre-specified equation or reference problem. Generated problems are:

1. Solvable: mathematically possible to solve
2. Accurate: solutions arrive at the correct answer
3. Appropriate: mathematically and contextually appropriate for a young learner

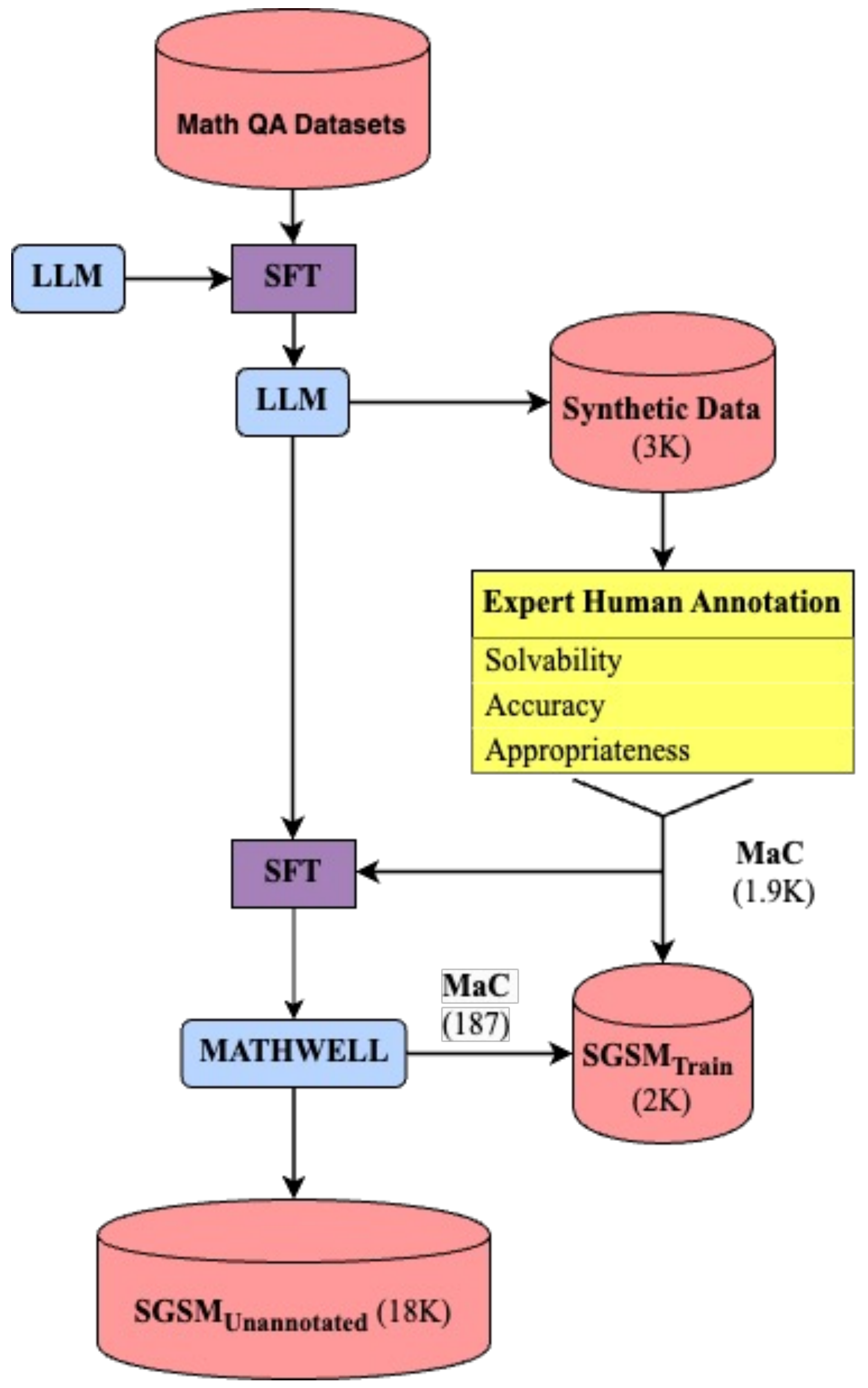
Data Generation, Expert Annotation, and Finetuning

We conduct two rounds of supervised finetuning (SFT) to create MATHWELL using:

1. Existing math QA data
2. Synthetic data annotated by domain experts with gold labels for meets all criteria (MaC), denoting questions that are solvable, accurate, and appropriate

We use MATHWELL to generate Synthetic Grade School Math (SGSM), consisting of two subsets:

1. $\text{SGSM}_{\text{Train}}$ with gold labels for MaC
2. $\text{SGSM}_{\text{Unannotated}}$ with executable code solutions but no labels

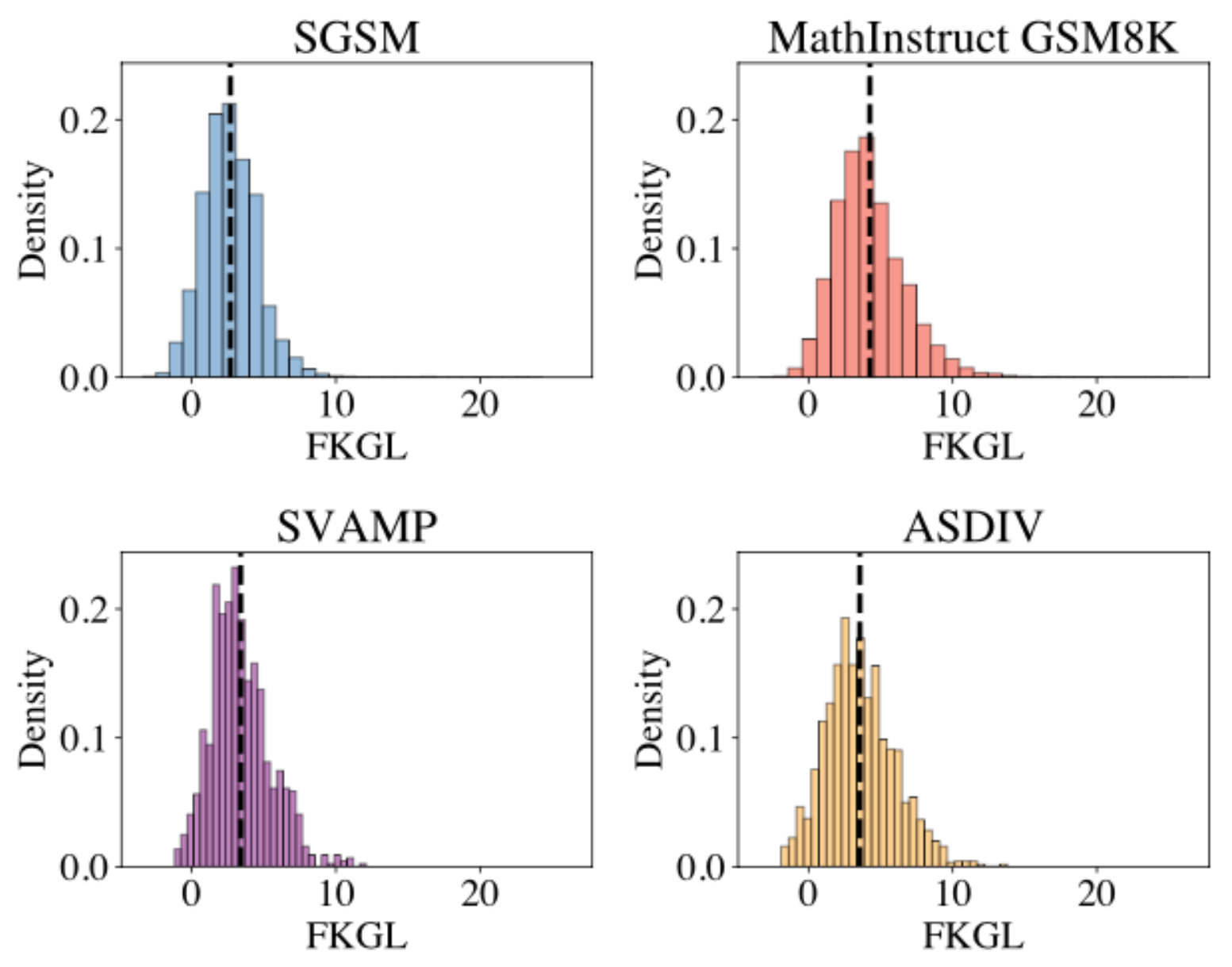


New Dataset: Synthetic Grade School Math (SGSM)

Dataset	N	Program of Thought (PoT)	Python Function	Appropriate Difficulty	Average Length (tokens)	Flesch-Kincaid Grade Level (FKGL)	New Dale-Chall (NDC) Readability	BERTScore F1
GSM-Hard	1,319	✓	✓	✗	72.9 (25.6)	4.21 (2.43)	8.20 (1.13)	84.0
MathInstruct GSM8K	6,403	✓	✗	✓	66.2 (23.9)	4.25 (2.48)	8.17 (1.13)	84.6
NumGLUE	12,403	✓	✗	✗	144.8 (136.5)	10.04 (6.99)	10.27 (1.51)	81.5
ASDIV	2,305	✗	✗	✓	45.1 (15.8)	3.56 (2.40)	7.85 (1.48)	85.5
SVAMP	1,000	✗	✗	✓	47.3 (11.7)	3.39 (2.07)	7.84 (1.09)	86.1
SGSM (Ours)	20,490	✓	✓	?	62.0 (15.0)	2.68 (1.97)	7.99 (1.26)	84.8
$\text{SGSM}_{\text{Train}}$	2,093	✓	✓	✓	57.2 (15.7)	2.50 (1.76)	8.12 (1.25)	85.2
$\text{SGSM}_{\text{Unannotated}}$	18,397	✓	✓	?	62.5 (14.8)	2.70 (1.99)	7.97 (1.26)	84.9

Advantages of SGSM:

1. Larger
2. Context-free
3. Mathematically appropriate
4. More readable
5. Human quality

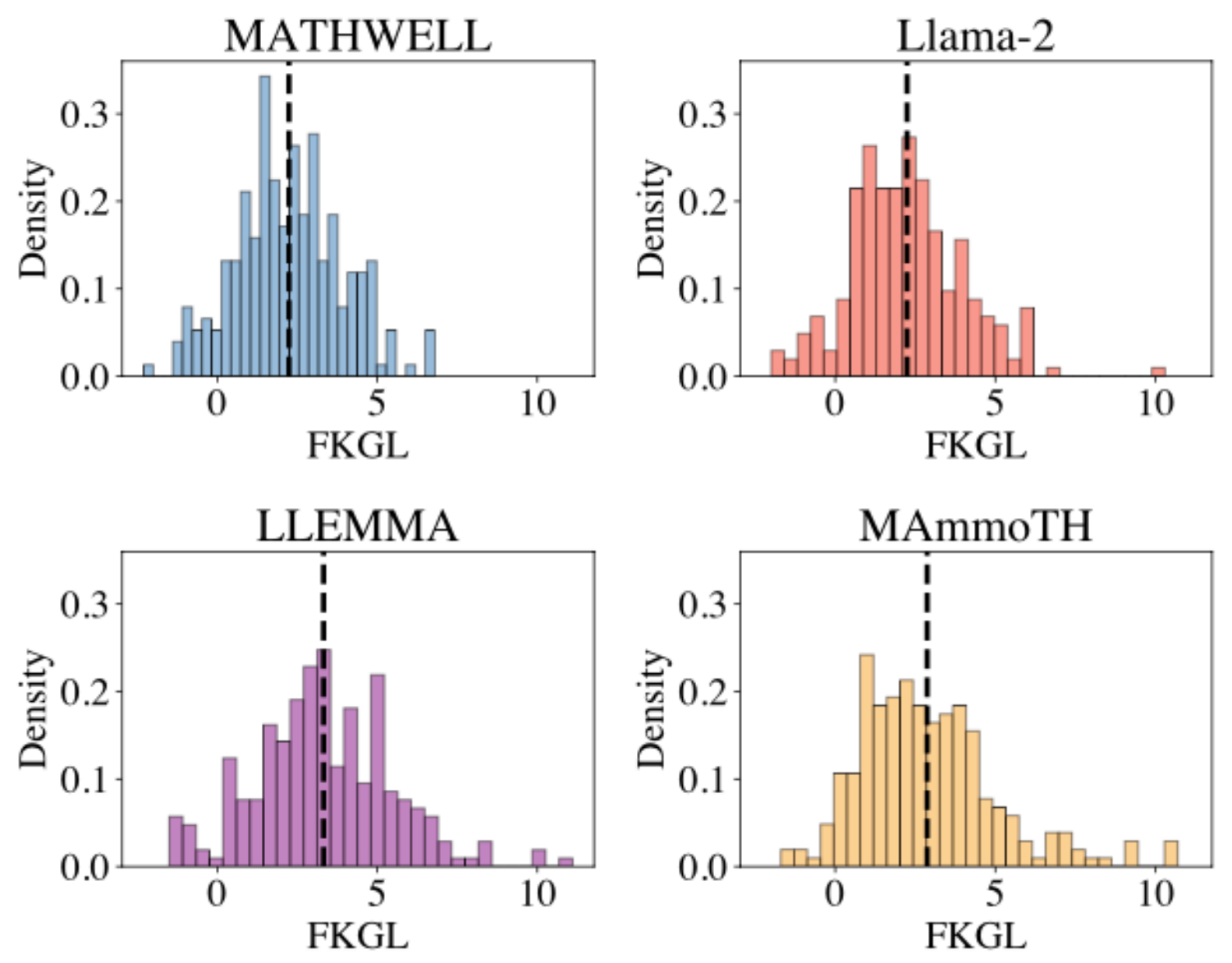


MATHWELL Outperforms Alternatives

Model	Solvability	Accuracy	Appropriateness	Meets all Criteria (MaC)	Topic Specificity	Executable Code	Executable Code/MaC
LLEMMA	48.8 (3.17)	63.9 (4.37)	41.8 (4.48)	15.2 (2.28)	94.8 (1.41)	24.3 (0.70)	3.70 (0.55)
MAmmoTH	86.8 (2.15)	94.9 (1.49)	67.7 (3.18)	56.8 (3.14)	97.6 (0.97)	6.90 (0.36)	3.91 (0.22)
Llama-2	84.0 (2.32)	89.5 (2.12)	81.0 (2.72)	62.4 (3.07)	99.2 (0.56)	55.4 (0.98)	34.6 (1.70)
MATHWELL	**89.2 (1.97)**	**96.9 (1.17)**	**86.5 (2.29)**	**74.8* (2.75)**	**99.6 (0.40)**	**66.4* (1.00)**	**49.6* (1.83)**

Model	PPL ↓	BertScore F1	GSM8K BERTScore F1	MaC Average Length	New Dale-Chall (NDC) Readability
LLEMMA	3.82 (0.10)	84.3	84.1	50.9 (2.89)	8.41 (0.09)
MAmmoTH	2.76 (0.03)	86.0	84.7	44.4 (1.15)	8.25 (0.08)
Llama-2	2.52 (0.03)	85.5	84.3	49.8 (1.19)	8.20 (0.07)
MATHWELL	**2.44 (0.03)**	85.5	84.2	**54.1 (0.97)**	8.23 (0.08)

MATHWELL outperforms alternatives in both human and automatic evaluations



MATHWELL outputs are:

1. More readable for the target age range (K-8)
2. Human quality
3. Highly customizable