ThinknCheck

Grounded Claim Verification with Compact, Reasoning-Driven, and Interpretable Models

Delip Rao, Feijiang Han, Chris Callison-Burch
University of Pennsylvania

Penn
University of Pennsylvania

TL;DR

- **Research Question:** How can we activate strong scientific claim verification in a small language model?
- **The Problem**
  - **Model:** Direct CoT prompting does not work reliably, and straightforward input–output fine-tuning gives only limited gains for 1B-scale models.
  - **Data:** Existing benchmarks mostly reward span matching, not multi-sentence or numerical reasoning.
- **Our Solution:**
  - **Data:** Create **LLMAggreFact-Think**, a 24.1K reasoning-augmented dataset; Construct **GSMClaims**, a new dataset for arithmetic claim verification.
  - **Model:** Build **ThinknCheck**, a 4-bit Gemma3-1B model fine-tuned to reason first, then decide; Build **ThinknCheck-Science**, a science-focused variant for real multi-step and numerical reasoning.
- **The Impact:** ThinknCheck-1B reaches 78.1 BAcc on LLMAggreFact (in-domain) and 64.7 BAcc on SciFact (out-of-domain), surpassing the larger SOTA MiniCheck-7B.

MOTIVATION

Claim verification sits at the core of **science, medicine, and public policy**: we rely on it to decide which findings to trust, which treatments to recommend, and which reports to act on. In these settings, we need systems that are not only **accurate** but also **grounded in evidence** and **easy to interpret.**

But today:

- Models: LLM-based verifiers are often large, closed, and expensive, limiting real-world deployment.
- Data: Common benchmarks over-reward span matching and under-test multi-sentence, scientific, and numerical reasoning.
- Small LMs: Naïve "make it think" methods (zero-shot CoT, simple GRPO rewards) can hurt 1B models, reinforcing lexical shortcuts instead of genuine reasoning.

👉 **Our work asks:** can a small model become a strong, interpretable claim verifier if we redesign both its training signals and evaluation benchmarks?

OUR METHOD

Key Intuition:

"Reason first, then decide" — even a 1B model can be a strong verifier if we train it to produce concise, task-specific rationales instead of raw labels or uncontrolled CoT.

Document

Atmospheric methane, Methane as a greenhouse gas: Methane in the Earth's atmosphere is a strong greenhouse gas with a global warming potential (GWP) 84 times greater than CO2 in a 20-year time frame; methane is not as persistent a gas as CO2 (assuming no change in carbon sequestration rates) and tails off to about GWP of 28 for a 100-year time frame. [...] Methane has a large effect but for a relatively brief period, having an estimated mean half-life of 9.1 years in the atmosphere, whereas carbon dioxide is currently given an estimated mean lifetime of over 100 years.

Claim

Carbon dioxide is a more potent greenhouse gas than methane.

→

Reasoning

The document specifically states that methane has a global warming potential (GWP) that is 84 times greater than that of carbon dioxide (CO2) in a 20-year time frame, and 28 times greater over a 100-year period. This indicates that methane is actually more potent than carbon dioxide as a greenhouse gas in both short and long terms. Therefore, the claim that carbon dioxide is a more potent greenhouse gas than methane contradicts the information provided in the document.

Verification

NO

Data & Benchmarks

1. **LLMAggreFact-Think** = LLMAggreFact + GPT-4o-mini reasoning chains + filter → 24.1K (document, claim, rationale, YES/NO) pairs
2. **GSMClaims** = GSM8K problems → GPT-4o rewrites as document + GPT-4o generates positive/negative claims + reasoning chains → arithmetic claim verification benchmark

Models

1. **ThinknCheck-1B =** 4-bit Gemma3-1B + SFT on LLMAggreFact-Think → <REASONING> concise rationale </REASONING> <SOLUTION> YES / NO </SOLUTION>
2. **ThinknCheck-Science** = ThinknCheck-1B + extra SFT on SciFact and GSMClaims → stronger scientific & numerical claim verification

KEY RESULTS

$$\text{BAcc} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

Model	BAcc
GPT-4 (zero-shot)	75.3
GPT-4o (zero-shot)	75.9
Claude-Sonnet-3.5 (zero-shot)	77.2
AlignScore (355M/fp16)	70.4
MiniCheck (7B/fp16)	77.4
Gemma3 (1B/fp4) ("base")	55.7
Gemma3 + CoT (1B/fp4)	51.4
ThinknCheck-nothink (1B/fp4)	57.5
ThinknCheck (1B/fp4)	**78.1**

Model	LLMAggre Fact	SciFact (dev)	GSM Claims
MiniCheck-7B	77.4	50.0	51.3
ThinknCheck-nothink-1B	57.5	21.7	49.4
ThinknCheck-1B	78.1	64.7	52.2
ThinknCheck-Science-1B	**79.2**	**66.4**	**61.0**

Method (Gemma3-1B, fp4)	BAcc
GRPO on Gemma3 ("base")	52.6
GRPO on Gemma3-SFT-400 ("warm start")	74.2

Claim: Momentum P ≈ 10.0 ±1.1 kg · m/s
Document: Mass M = 5.0 ±0.1 kg. Velocity V = 2.0 ±0.2 m/s.

MiniCheck-7B verification result:
DOES NOT SUPPORT

ThinknCheck-Science-1B verification result:
SUPPORTS

ThinknCheck-Science-1B reasoning:
The momentum is calculated by multiplying the mass and the velocity: 5.0 kg ×2.0 m/s = 10.0 kg · m/s. The relative uncertainty in the mass is 0.1/5.0 = 2% and in the velocity is 0.2/2.0 = 10%. When uncertainties are combined for multiplication, they add in quadrature: √((2%)² + (10%)²) = √(0.0004 + 0.01) = √(0.0104) ≈ 10.2%. This results in an uncertainty of about 10.2% ×10.0 kg · m/s ≈ 1.0 kg · m/s. A better calculation using maximum and minimum values yields P_{max} = 5.1 kg ×2.2 m/s = 11.22 kg · m/s and P_{min} = 4.9 kg ×1.8 m/s = 8.82 kg · m/s. The range between P_{max} and P_{min} is 11.22 8.82 = 2.4 kg · m/s, implying an uncertainty of approximately ±1.2 kg · m/s. Since both approaches yield an uncertainty around ±1.0 to ±1.2 kg · m/s, the claim of 10.0 ±1.1 kg · m/s is consistent with the uncertainties derived from the document.

TAKE AWAYS

1. **Small but smart:** A 1B, 4-bit verifier with supervised reasoning can match or beat 7B and closed models, while being cheap and deployable.
2. **Reasoning matters:** Explicit, supervised reasoning is far more effective than naïve CoT or simple RL at the 1B scale.
3. **Benchmarks must evolve:** Current benchmarks under-test multi-step and numerical reasoning—we need better tasks and rationales to assess claim verification truly.