

Machine-Verifiable Item Models for Geometry Diagram Generation

Mei Chen · Gordon Stein · AIME-Con 2026





Testing runs on fresh items

Every test needs a steady supply of new items, each with its answer key.

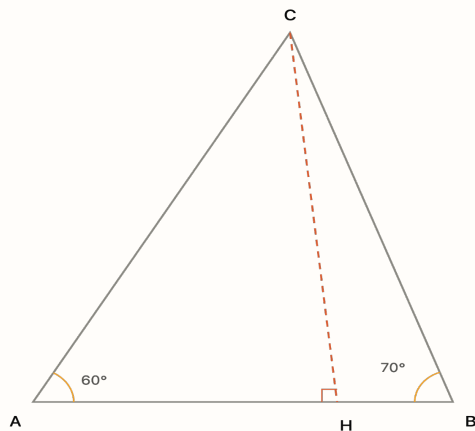
Automatic item generation is the field's answer. It has gone from templates to language models to agents. Almost all of it is text.

Geometry diagrams are still drawn by hand and reviewed by hand.

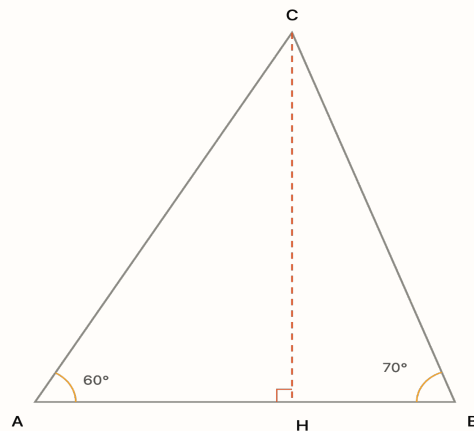


Ask a language model to draw

“Draw an acute triangle ABC with $\angle A = 60^\circ$ and $\angle B = 70^\circ$, then draw the altitude from C , meeting AB at H .”



Fails the checks. The foot H and the angles are not what was asked.



Passes the checks.



Why the existing answers fall short

Fixed templates render specific diagrams reliably. Each is hand-built. None generalizes.

Geometry benchmarks test whether a model can read a diagram, not whether it can draw one.

Image benchmarks ask a vision model to judge requirements that are fully deterministic.

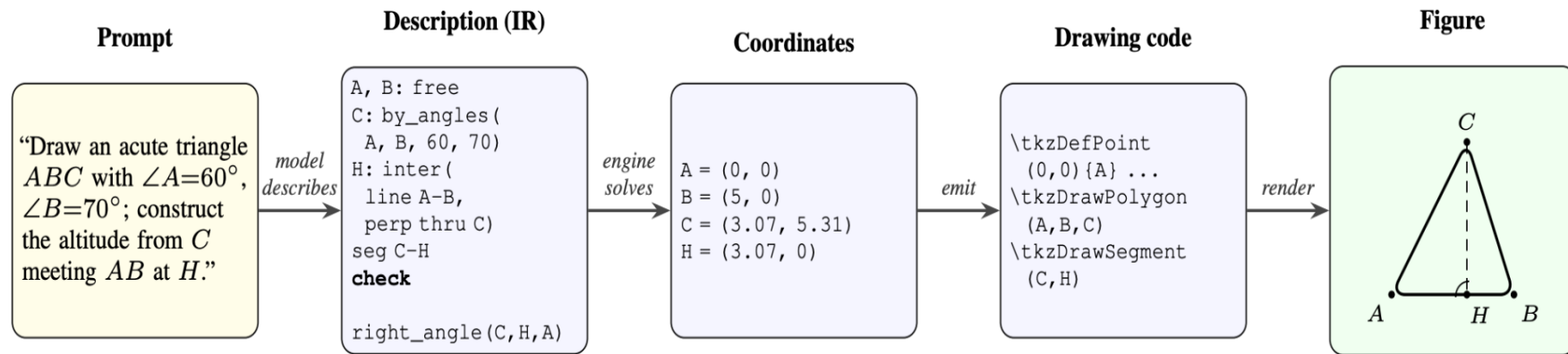


The language model supplies *flexibility*. The geometry engine supplies *validity*. The item model ties generation to *scoring*.

The model describes the construction. It never places a point.



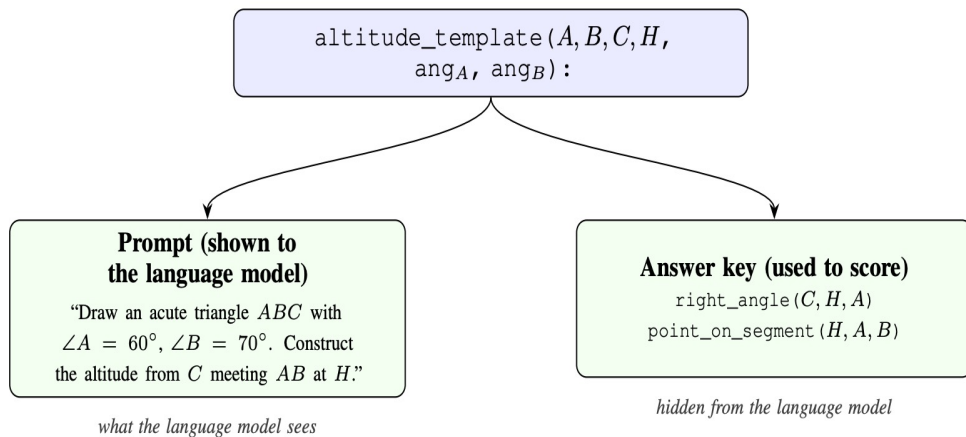
How the model draws a diagram



Describe, solve, emit, render. The engine computes every coordinate, so the drawing cannot contradict the description.



The question and its key are written together



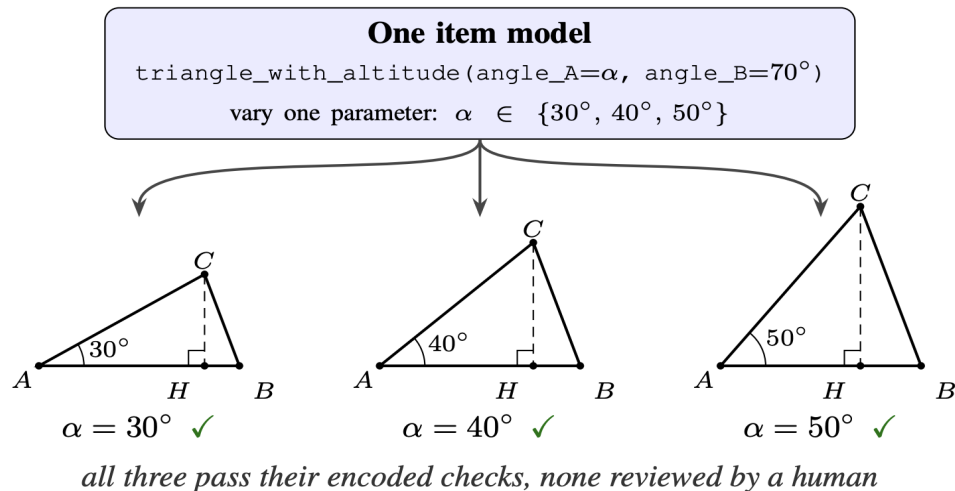
One function writes both, from the same variables. They cannot drift apart.

Fifteen checks, each arithmetic on the coordinates. A right angle is a dot product, not an opinion.

The model never sees the key.



One item model, a family of items



Vary one parameter and every sibling arrives with its own key.

All three pass their checks. None was reviewed by a person.



The benchmark

Every drawing is scored by the same checker, so the two modes and the six models are measured with one instrument.

600

prompts from 30 item models in three tiers

6

language models from three vendors

2

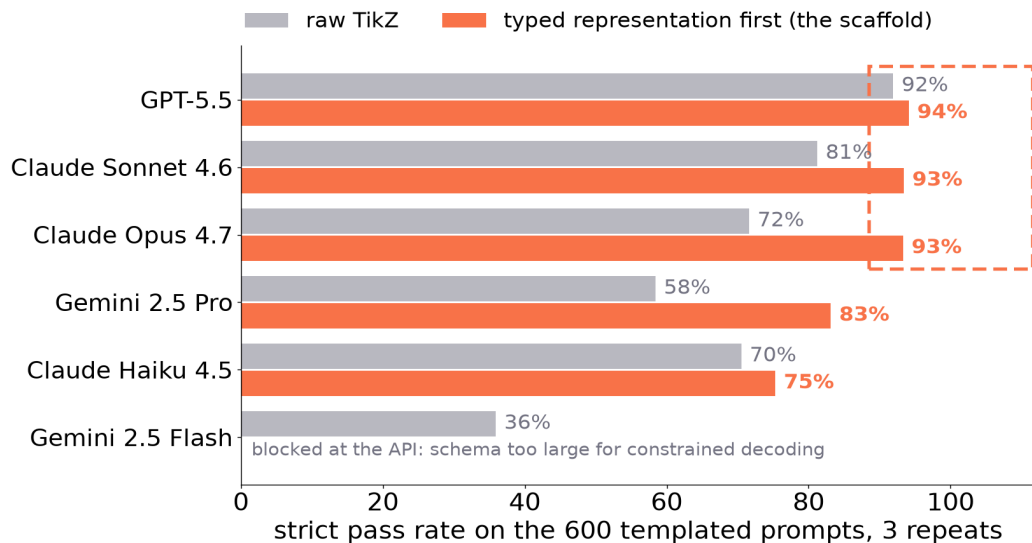
modes: by-hand drawing code, or a description first

3

attempts each, every output scored by the checker



Structured mode helped every model that could run it



Gains of 2 to 25 points. The weaker a model is at drawing code, the more it gains.

Graders confirmed 75% of by-hand passes and 87% of structured passes. The advantage holds after correction.

Gemini Flash: blocked at the API, not by the model.



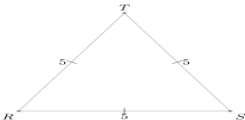
How we checked the checker

GeoGen-rating rater_a Item 1 / 200 ✓ 0 - 0 ✗ 0 200 left [next unanswered](#) saved Help

Workflow: read prompt → eyeball diagram → answer Q1 blind → Q2 unlocks per-predicate. Press [?](#) for shortcuts.

DIAGRAM

tpl-t1-eq-RST-s5 · claude-opus-4-7 · raw_code



Ruler Protractor

PROMPT

T1 raw_code

TASK

Create a diagram showing an equilateral triangle RST where every side has length 5. Label all three vertices.

Item metadata + TikZ source

RUBRIC

Q1 — Overall verdict

Does the diagram correctly satisfy every claim in the prompt? (auto-verdict hidden until you answer) · Yes = fully satisfied · Partial = some clauses violated · No = major violation / unrelated

☒ Yes 1 ☐ Partial 2 ☐ No 3

Q2 — Per-property agreement

For each property the verifier checked, do you agree with its verdict?

[Submit Q1 to unlock per-property questions \(blinding\)](#)

+ Add notes

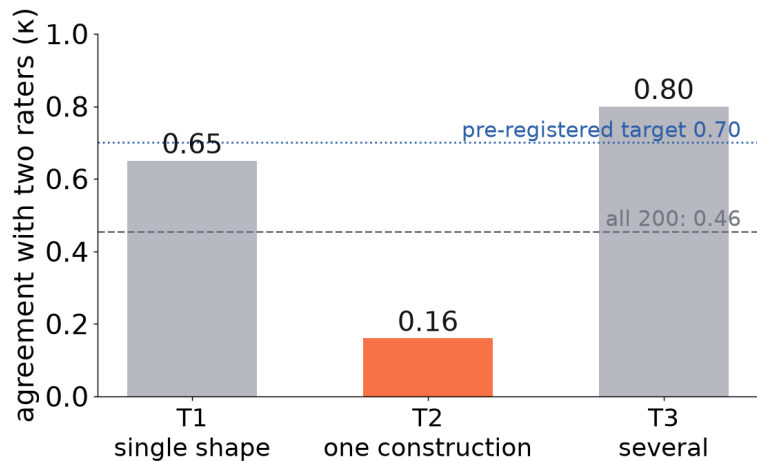
Prev Next

0/200 Q1 done

A grader sees the prompt and the diagram, nothing else. The blind verdict comes first. The per-property questions unlock only after it.



Does the scorer agree with people?



200 diagrams, two graders blind to the scorer. Target $\kappa \geq 0.70$.

Result 0.46; the graders agreed with each other at 0.70.

The disagreement runs one way: the scorer never failed a diagram the graders passed.

Per requirement it matched them on 98% of 864 ratings. The leak sits in one tier, where the checklist pins the construction but not the numbers.



What the checks cannot see

The checks read geometry, not presentation. A vision judge sees how a picture looks and cannot check its geometry. The two are complementary.

A checklist can leave a requirement out, and an item model takes work to write.

Fifteen checks do not cover every property a prompt might ask for.



Why this matters for measurement

Sweep one parameter and you get a family of siblings with built-in keys.

Quality control moves from reviewing each item to reviewing the item model.

Open question: are the siblings true isomorphs? Difficulty needs field testing.



Structured generation improves diagram quality and yields item models with the key built in.

Scoring needs stronger agreement with human graders before operational use.

Next: 3D and coordinate graphs, and constructions composed from steps so the checks come with them.

mei.chen@renphil.org · gordon.stein@renphil.org