# Comparing human perception, computer-algebra, and generative AI approaches for ranking elementary geometry statements

Z. Kovács[1] T. Recio[2] P. Tolmos[3] M. P. Vélez[2]

[1]Priv. Univ. of Educ., Dioc. Linz [2]Univ. Nebrija, Madrid [3]Univ. Rey Juan Carlos, Madrid

SCML-2026, July 6-8, 2026, Hagenberg City, Austria

# Is this a difficult problem?

Let $\Omega$ and $\Gamma$ be circles with centres $M$ and $N$, respectively, such that the radius of $\Omega$ is less than the radius of $\Gamma$. Suppose circles $\Omega$ and $\Gamma$ intersect at two distinct points $A$ and $B$. Line $MN$ intersects $\Omega$ at $C$ and $\Gamma$ at $D$, such that points $C$, $M$ and $N$ lie on the line in that order. Let $P$ be the circumcentre of triangle $ACD$. Line $AP$ intersects $\Omega$ again at $E \neq A$. Line $AP$ intersects $\Gamma$ again at $F \neq A$. Let $H$ be the orthocentre of triangle $PMN$.

Prove that the line through $H$ parallel to $AP$ is tangent to the circumcircle of triangle $BEF$. (The *orthocentre* of a triangle is the point of intersection of its altitudes.)

# Is this a difficult problem?

Let $\Omega$ and $\Gamma$ be circles with centres $M$ and $N$, respectively, such that the radius of $\Omega$ is less than the radius of $\Gamma$. Suppose circles $\Omega$ and $\Gamma$ intersect at two distinct points $A$ and $B$. Line $MN$ intersects $\Omega$ at $C$ and $\Gamma$ at $D$, such that points $C$, $M$ and $N$ lie on the line in that order. Let $P$ be the circumcentre of triangle $ACD$. Line $AP$ intersects $\Omega$ again at $E \neq A$. Line $AP$ intersects $\Gamma$ again at $F \neq A$. Let $H$ be the orthocentre of triangle $PMN$.

Prove that the line through $H$ parallel to $AP$ is tangent to the circumcircle of triangle $BEF$. (The *orthocentre* of a triangle is the point of intersection of its altitudes.)

Who cares? ¯_(ツ)_/¯

# Is this a difficult problem?

Let $\Omega$ and $\Gamma$ be circles with centres $M$ and $N$, respectively, such that the radius of $\Omega$ is less than the radius of $\Gamma$. Suppose circles $\Omega$ and $\Gamma$ intersect at two distinct points $A$ and $B$. Line $MN$ intersects $\Omega$ at $C$ and $\Gamma$ at $D$, such that points $C$, $M$ and $N$ lie on the line in that order. Let $P$ be the circumcentre of triangle $ACD$. Line $AP$ intersects $\Omega$ again at $E \neq A$. Line $AP$ intersects $\Gamma$ again at $F \neq A$. Let $H$ be the orthocentre of triangle $PMN$.

Prove that the line through $H$ parallel to $AP$ is tangent to the circumcircle of triangle $BEF$. (The *orthocentre* of a triangle is the point of intersection of its altitudes.)

Who cares? ¯_(ツ)_/¯

This is Problem 2 of the International Mathematical Olympiad 2025. :-)

# Problem types of the IMO

| *Year* | *Country* | *Problems* | *Proofs problems in Euclidean geometry* |
|---|---|---|---|
| 2014 | South Africa | 6 | 2 |
| 2015 | Thailand | 6 | 2 |
| 2016 | Hong Kong | 6 | 1 |
| 2017 | Brazil | 6 | 1 |
| 2018 | Romania | 6 | 2 |
| 2019 | UK | 6 | 2 |
| 2020 | Russia | 6 | 1 |
| 2021 | Russia | 6 | 2 |
| 2022 | Norway | 6 | 1 |
| 2023 | Japan | 6 | 2 |
| 2024 | UK | 6 | 1 |
| 2025 | Australia | 6 | 1 |
| 2026 | China | ? | ? |

# Aren’t we banging on open doors?

## AI can already win a gold medal at IMO since 2025!

**Computer Science > Artificial Intelligence**

*[Submitted on 21 Jul 2025 (v1), last revised 30 Sep 2025 (this version, v4)]*

# Winning Gold at IMO 2025 with a Model-Agnostic Verification-and-Refinement Pipeline

Yichen Huang, Lin F. Yang

The International Mathematical Olympiad (IMO) is widely regarded as the world championship of high-school mathematics. IMO problems are renowned for their difficulty and novelty, demanding deep insight, creativity, and rigor. Although large language models perform well on many mathematical benchmarks, they often struggle with Olympiad-level problems. Using carefully designed prompts, we construct a model-agnostic, verification-and-refinement pipeline. We demonstrate its effectiveness on the recent IMO 2025, avoiding data contamination for models released before the competition. Equipped with any of the three leading models -- Gemini 2.5 Pro, Grok-4, or GPT-5 -- our pipeline correctly solved 5 out of the 6 problems (≈85.7% accuracy). This is in sharp contrast to their baseline accuracies: 31.6% (Gemini 2.5 Pro), 21.4% (Grok-4), and 38.1% (GPT-5), obtained by selecting the best of 32 candidate solutions. The substantial improvement underscores that the path to advanced AI reasoning requires not only developing more powerful base models but also designing effective methodologies to harness their full potential for complex tasks.

Subjects: **Artificial Intelligence (cs.AI)**

Cite as: arXiv:2507.15855 **[cs.AI]**

# Yes, but. . .

All that “large” AI systems (with or without verification/refinement) can do is

- a black box for most users (perhaps nobody really knows how it works, why it actually works)
- non-deterministic for most users (you never get exactly the same answer again)
- unreliable (sometimes it’s wonderful, sometimes it’s okay, but sometimes you get garbage)
- dependent on available resources (what if there is not enough RAM, not enough nodes in the supercomputer?)
- dependent on software and version (what if a new version does not support your favorite topic anymore?)

# Yes, but...

All that "large" AI systems (with or without verification/refinement) can do is

- a black box for most users (perhaps nobody really knows how it works, why it actually works)
- non-deterministic for most users (you never get exactly the same answer again)
- unreliable (sometimes it's wonderful, sometimes it's okay, but sometimes you get garbage)
- dependent on available resources (what if there is not enough RAM, not enough nodes in the supercomputer?)
- dependent on software and version (what if a new version does not support your favorite topic anymore?)

$\rightarrow$ *Proving* anyway requires
a reproducible and affordable workflow!

# How to identify. . .

## . . .suitable problems for contests, test papers, exams?

- Manual check (by humans) – biased
- Machine check (with algorithms) – unbiased*

---

* *less* biased

# How to identify…

…suitable problems for contests, test papers, exams?

- Manual check (by humans) – biased
- Machine check (with algorithms) – unbiased*

→ Algorithms must implement a ranking standard!

(It has to be *invented* — by humans!)

* *less* biased

# Software mentioned in this presentation

- *GeoGebra*, a free dynamic mathematics software with educational concept in mind (maintained in Linz)

# Software mentioned in this presentation

- *GeoGebra*, a free dynamic mathematics software with educational concept in mind (maintained in Linz)
- *GeoGebra Discovery*, a fork of GeoGebra, with focus on automated geometry reasoning (maintained in Linz and Madrid, by the authors)

# Software mentioned in this presentation

- *GeoGebra*, a free dynamic mathematics software with educational concept in mind (maintained in Linz)
- *GeoGebra Discovery*, a fork of GeoGebra, with focus on automated geometry reasoning (maintained in Linz and Madrid, by the authors)
- AI chatbots with partial automated reasoning capabilities, here in particular *ChatGPT* (developed by OpenAI)

# Software mentioned in this presentation

- *GeoGebra*, a free dynamic mathematics software with educational concept in mind (maintained in Linz)
- *GeoGebra Discovery*, a fork of GeoGebra, with focus on automated geometry reasoning (maintained in Linz and Madrid, by the authors)
- AI chatbots with partial automated reasoning capabilities, here in particular *ChatGPT* (developed by OpenAI)
- Computer-algebra systems like *Mathematica* (by Wolfram), *Maple* (by Maplesoft) and *Giac* (by Univ. of Grenoble II)

# Software mentioned in this presentation

- *GeoGebra*, a free dynamic mathematics software with educational concept in mind (maintained in Linz)
- *GeoGebra Discovery*, a fork of GeoGebra, with focus on automated geometry reasoning (maintained in Linz and Madrid, by the authors)
- AI chatbots with partial automated reasoning capabilities, here in particular *ChatGPT* (developed by OpenAI)
- Computer-algebra systems like *Mathematica* (by Wolfram), *Maple* (by Maplesoft) and *Giac* (by Univ. of Grenoble II)
- *Java Geometry Expert*, a free dynamic geometry software with educational concept and focus on automated proofs (maintained by ADG Foundation)

# Focused features of the mentioned software tools

*GeoGebra Discovery*, `ShowProof` and `DescribeStatement` commands

- User input:
  - a step-by-step construction created using intuitive GeoGebra tools (widely used in many high schools) — the **hypotheses**
  - a geometric statement (e.g. parallelism, perpendicularity, an algebraic formula) to prove — the **thesis**
- Output:
  - A certified step-by-step algebraic proof which assumes some non-degeneracy conditions as extra hypotheses
  - Complexity grade of planar geometry theorems
  - Exportable proof to Mathematica, Maple, and Giac formats for further analysis
  - Exportable statement in textual format (in various human languages)

# Focused features of the mentioned software tools

## *Java Geometry Expert*, Geometry Database Deduction Method

- User input:
  - a hypotheses + thesis in GeoGebra format
- Output:
  - A syntetic proof
  - Complexity grade measured via proof graph properties (number of vertices, edges)

# Focused features of the mentioned software tools

## *ChatGPT*

- User input:
  - hypotheses + thesis in JGEX format
- Output:
  - A textual proof (unreliable, non-deterministic, black box)
  - Complexity grade measured via properties of the text output (e.g. character length)

# Focused features of the mentioned software tools

*ChatGPT*

- Pros:
  - Appropriate ideas (good matches with textbooks methods)
  - Correct numerical and symbolic computations (reliable connections to mathematics software)
- Cons:
  - The whole picture may be incomplete or wrong
  - Personalized context is difficult to set up ("I am an Austrian student with a strong algebraic background but poor geometric understanding with a limitation to know only the Pythagorean theorem, and I know nothing about projective planes, etc., can you still help me proving Desargues's theorem?")
  - Circular proofs and other "surprises"

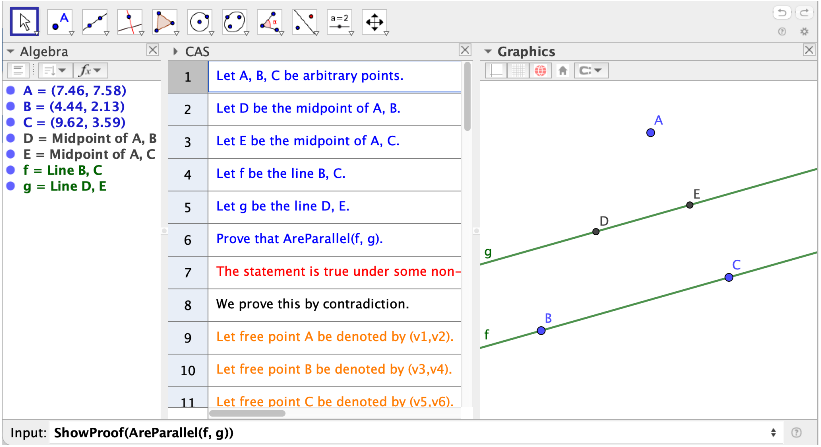
# Focused features of the mentioned software tools

*Mathematica*, *Maple* and *Giac*

- User input:
  - An algebraic computation
    $f_1 \cdot h_1 + \ldots + f_r \cdot h_r + f_{r+1} \cdot (h_{r+1} \cdot z_1 - 1) + f_{r+2} \cdot (t \cdot z_2 - 1) = 1$
    that corresponds to a proof by contradiction, written in the input language of the software tool (by )
- Output:
  - $1 = 0$ if the proof is certified successfully

# Example 1: The midline theorem

GeoGebra Discovery, `ShowProof` command



# Example 1: The midline theorem

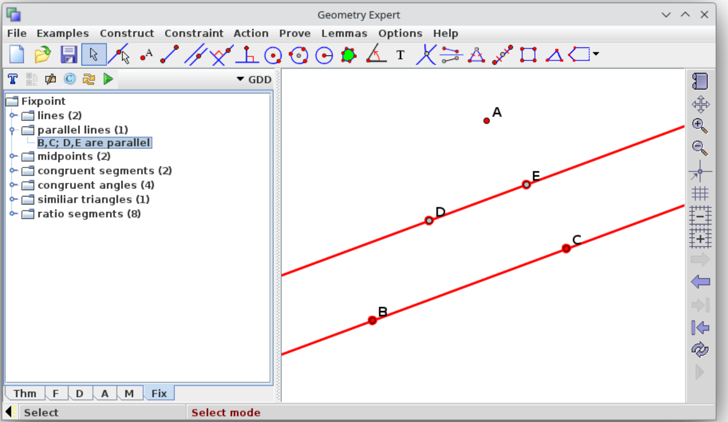
## GeoGebra Discovery, `ShowProof` command

CAS

| | |
|---|---|
| 20 | Thesis: AreParallel(f, g), in algebraic form: |
| 21 | T1:=-v4*v7+v6*v7+v3*v8-v5*v8+v4*v9-v6*v9-v3*v10+v5*v10=0<br>→ **T1 : $-$v10 v3 $+$ v10 v5 $+$ v3 v8 $-$ v4 v7 $+$ v4 v9 $-$ v5 v8 $+$ v6 v7 $-$ v6 v9 $=$ 0** |
| 22 | Thesis reductio ad absurdum (denied statement): |
| 23 | v11: dummy variable to express negation |
| 24 | (T1*v11-1)=0<br>→ **v11 ($-$v10 v3 $+$ v10 v5 $+$ v3 v8 $-$ v4 v7 $+$ v4 v9 $-$ v5 v8 $+$ v6 v7 $-$ v6 v9) $-$ 1 $=$ 0** |
| 25 | e5:=-1-v4*v7*v11+v6*v7*v11+v3*v8*v11-v5*v8*v11+v4*v9*v11-v6*v9*v11-v3*v10*v11+v5*v10*v11=0<br>→ **e5 : $-$v10 v11 v3 $+$ v10 v11 v5 $+$ v11 v3 v8 $-$ v11 v4 v7 $+$ v11 v4 v9 $-$ v11 v5 v8 $+$ v11 v6 v7 $-$** |
| 26 | Without loss of generality, some coordinates can be fixed: |
| 27 | {v4=1,v3=0,v2=0,v1=0}<br>→ **{v4 $=$ 1, v3 $=$ 0, v2 $=$ 0, v1 $=$ 0}** |
| 28 | The statement requires some conditions: |
| 29 | • A and B are not equal |
| 30 | All hypotheses and the negated thesis after |
| 31 | s1:2*v7=0<br>→ **s1 : 2 v7 $=$ 0** |
| 32 | s2:-1+2*v8=0<br>→ **s2 : 2 v8 $-$ 1 $=$ 0** |
| 33 | s3:-v5+2*v9=0<br>→ **s3 : $-$v5 $+$ 2 v9 $=$ 0** |
| 34 | s4:-v6+2*v10=0<br>→ **s4 : 2 v10 $-$ v6 $=$ 0** |
| 35 | s5:-1-v7*v11+v6*v7*v11-v5*v8*v11+v9*v11-v6*v9*v11+v5*v10*v11=0<br>→ **s5 : v10 v11 v5 $-$ v11 v5 v8 $+$ v11 v6 v7 $-$ v11 v6 v9 $-$ v11 v7 $+$ v11 v9 $-$ 1 $=$ 0** |
| 36 | Now we consider the following equation: |
| 37 | s1*(v10*v11-1/2*v11)+s2*(-v11*v9)+s3*(-v10*v11+v11*v8)+s4*(-v11*v7+v11*v9)+s5*(-1)<br>→ **1 $=$ 0** |
| 38 | **Contradiction! This proves the original statement.** |
| 39 | The statement has a difficulty of degree 2. |

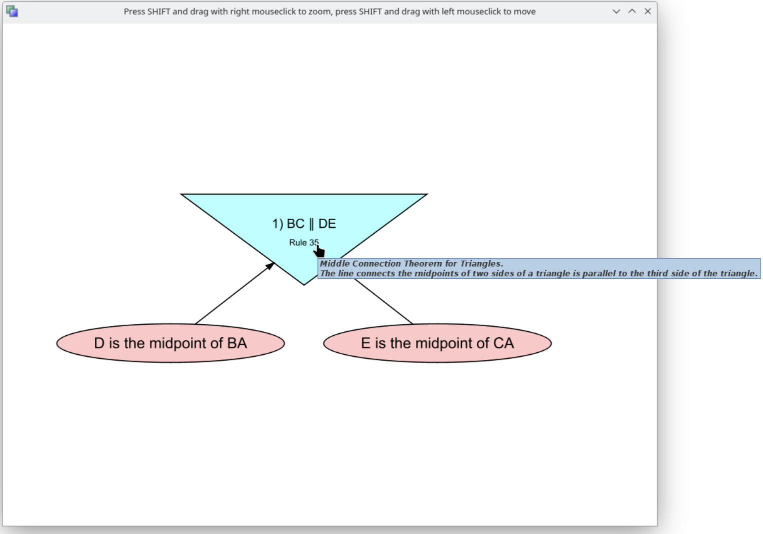
# Example 1: The midline theorem

## Java Geometry Expert, Geometric Deductive Database Method



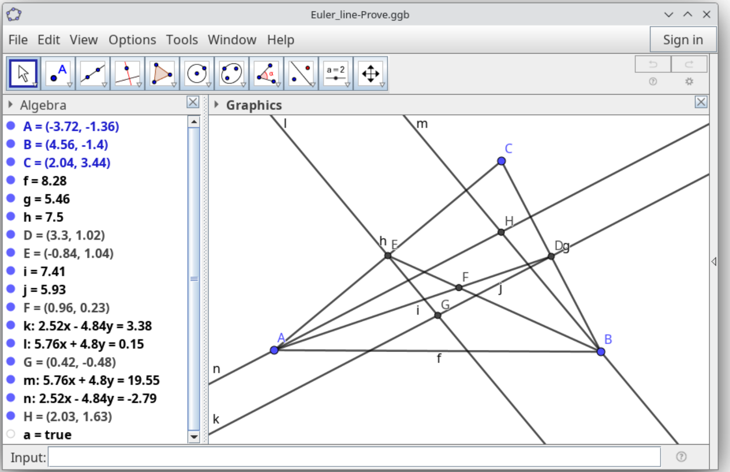
# Example 1: The midline theorem

## Java Geometry Expert, Geometric Deductive Database Method



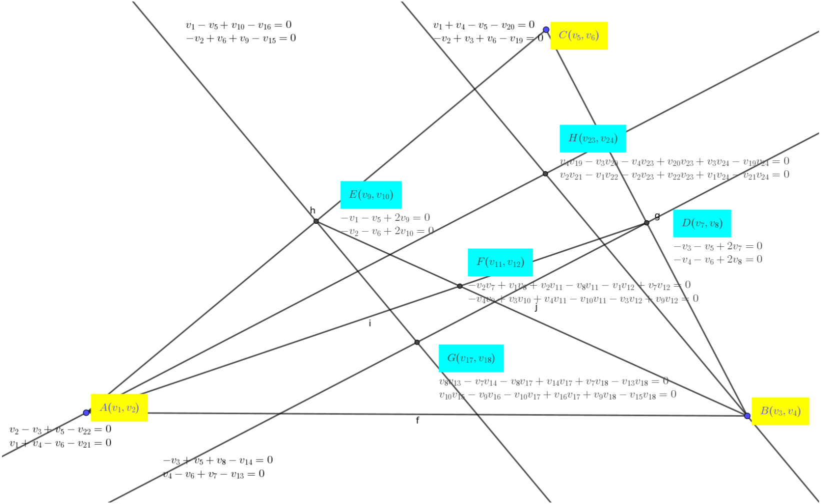
# Example 2: Euler's line

GeoGebra Discovery, `ShowProof` command



# Example 2: Euler's line

GeoGebra Discovery, `ShowProof` command



# Example 2: Euler's line

## GeoGebra Discovery, `ShowProof` command

# Example 2: Euler's line

## GeoGebra Discovery, `ShowProof` command

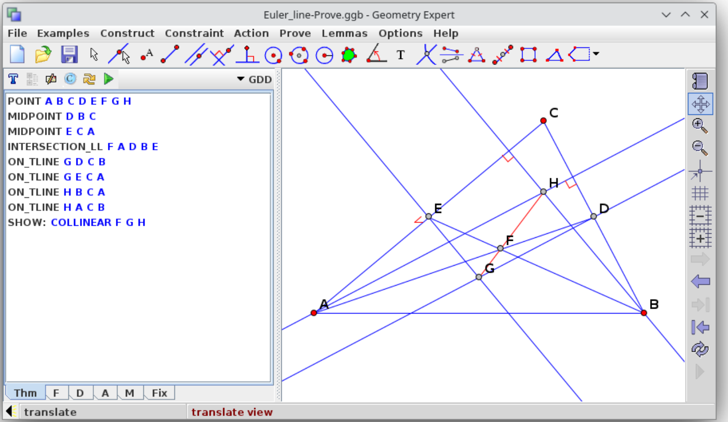
Euler_line-Prove.m

home > kovzol > Euler_line-Prove.m

```
s18:=(v22 * v23) - (v21 * v24) == 0;
s19:=-1 - ((v12 * v17) * v25) + ((v11 * v18) * v25) + ((v12 * v23) * v25) - ((v18 * v23) * v25) - ((v11 * v24) * v25) + ((v17 * v24) * v25) == 0;
(* Now we consider the following equation: *)
(s1 * (((((-1 / 3 * v25) * v26) * v12) * v17) * v6) + (((((1 / 3 * v25) * v26) * v12) * v23) * v6) + ((((1 / 3 * v25) * v26) * v14) * v6^(2)) - (((1 / 6 * v25) * v26) *
v6^(3)) + ((((2 / 3 * v25) * v26) * v12) * v17) - ((((2 / 3 * v25) * v26) * v12) * v23) - ((((1 / 6 * v25) * v26) * v14) * v6) - (((1 / 12 * v25) * v26) * v6^(2)) + (((1
/ 3 * v25) * v12) * v18) + (((1 / 2 * v25) * v14) * v23) - (((1 / 3 * v25) * v12) * v24) + (((1 / 12 * v25) * v26) * v6) - (((1 / 4 * v25) * v23) * v6) - (((1 / 6 * v25)
* v14) * v5) + (((1 / 12 * v25) * v6) * v5) - ((1 / 4 * v25) * v23) + ((1 / 12 * v25) * v5))) + (s2 * ((((((1 / 3 * v25) * v26) * v11) * v17) * v6) - (((((1 / 3 * v25) *
v26) * v11) * v23) * v6) - ((((1 / 3 * v25) * v26) * v13) * v6^(2)) + ((((1 / 6 * v25) * v26) * v6^(2)) * v5) - ((((2 / 3 * v25) * v26) * v11) * v17) + ((((2 / 3 * v25)
* v26) * v11) * v23) + ((((1 / 6 * v25) * v26) * v13) * v6) - ((((1 / 12 * v25) * v26) * v6) * v5) - (((1 / 3 * v25) * v11) * v18) - (((1 / 2 * v25) * v13) * v23) + (((1
/ 3 * v25) * v11) * v24) + (((1 / 6 * v25) * v13) * v5) + (((1 / 4 * v25) * v23) * v5) - ((1 / 12 * v25) * v5^(2)))) + (s3 * ((((((1 / 3 * v25) * v26) * v12) * v17) *
v6) - (((((1 / 3 * v25) * v26) * v12) * v23) * v6) + ((((1 / 3 * v25) * v26) * v16) * v6^(2)) - (((1 / 6 * v25) * v26) * v6^(3)) + ((((1 / 3 * v25) * v26) * v12) * v17)
- ((((1 / 3 * v25) * v26) * v12) * v23) - ((((1 / 2 * v25) * v26) * v16) * v6) - ((((1 / 3 * v25) * v26) * v17) * v6) + ((((1 / 3 * v25) * v26) * v23) * v6) + (((1 / 4 *
v25) * v26) * v6^(2)) + (((1 / 6 * v25) * v26) * v16) - (((1 / 3 * v25) * v26) * v17) - (((1 / 3 * v25) * v12) * v18) + (((1 / 3 * v25) * v26) * v23) + (((1 / 2 * v25) *
v16) * v23) + (((1 / 3 * v25) * v12) * v24) - (((1 / 12 * v25) * v26) * v6) - (((1 / 4 * v25) * v23) * v6) - (((1 / 6 * v25) * v16) * v5) + (((1 / 12 * v25) * v6) * v5)
+ ((1 / 3 * v25) * v18) - ((1 / 3 * v25) * v24))) + (s4 * ((((((-1 / 3 * v25) * v26) * v11) * v17) * v6) + (((((1 / 3 * v25) * v26) * v11) * v23) * v6) - ((((1 / 3 *
v25) * v26) * v15) * v6^(2)) + ((((1 / 6 * v25) * v26) * v6^(2)) * v5) - ((((1 / 3 * v25) * v26) * v11) * v17) + ((((1 / 3 * v25) * v26) * v11) * v23) + ((((1 / 2 * v25)
* v26) * v15) * v6) - ((((1 / 4 * v25) * v26) * v6) * v5) - (((1 / 6 * v25) * v26) * v15) + (((1 / 3 * v25) * v11) * v18) - (((1 / 2 * v25) * v15) * v23) - (((1 / 3 *
v25) * v11) * v24) + (((1 / 12 * v25) * v26) * v5) + (((1 / 6 * v25) * v15) * v5) + (((1 / 4 * v25) * v23) * v5) - ((1 / 12 * v25) * v5^(2)))) + (s5 * (((((2 / 3 * v25)
* v26) * v17) * v6) - ((((2 / 3 * v25) * v26) * v23) * v6) - (((4 / 3 * v25) * v26) * v17) + (((4 / 3 * v25) * v26) * v23) - ((2 / 3 * v25) * v18) + ((2 / 3 * v25) *
v24))) + (s6 * (((((-2 / 3 * v25) * v26) * v17) * v6) + ((((2 / 3 * v25) * v26) * v23) * v6) - (((2 / 3 * v25) * v26) * v17) + (((2 / 3 * v25) * v26) * v23) + ((2 / 3 *
v25) * v18) - ((2 / 3 * v25) * v24))) + (s7 * (((((2 / 3 * v25) * v26) * v17) * v6^(2)) - ((((1 / 3 * v25) * v26) * v6^(2)) * v5) - ((((1 / 3 * v25) * v26) * v17) * v6)
+ ((((1 / 6 * v25) * v26) * v6) * v5) + ((v25 * v17) * v23) - (((1 / 3 * v25) * v17) * v5) - (((1 / 2 * v25) * v23) * v5) + ((1 / 6 * v25) * v5^(2)))) + (s8 * (((((-2 /
3 * v25) * v26) * v18) * v6^(2)) + (((1 / 3 * v25) * v26) * v6^(3)) + ((((1 / 3 * v25) * v26) * v18) * v6) + (((1 / 6 * v25) * v26) * v6^(2)) - ((v25 * v18) * v23) -
(((1 / 6 * v25) * v26) * v6) + (((1 / 2 * v25) * v23) * v6) + (((1 / 3 * v25) * v18) * v5) - (((1 / 6 * v25) * v6) * v5) + ((1 / 2 * v25) * v23) - ((1 / 6 * v25) * v5)))
+ (s9 * (((((2 / 3 * v25) * v26) * v17) * v6^(2)) - ((((1 / 3 * v25) * v26) * v6^(2)) * v5) - (((v25 * v26) * v17) * v6) + ((((1 / 2 * v25) * v26) * v6) * v5) + (((1 / 3
* v25) * v26) * v17) + ((v25 * v17) * v23) - (((1 / 6 * v25) * v26) * v5) - (((1 / 3 * v25) * v17) * v5) - (((1 / 2 * v25) * v23) * v5) + ((1 / 6 * v25) * v5^(2)))) +
(s10 * (((((-2 / 3 * v25) * v26) * v18) * v6^(2)) + (((1 / 3 * v25) * v26) * v6^(3)) + (((v25 * v26) * v18) * v6) - (((1 / 2 * v25) * v26) * v6^(2)) - (((1 / 3 * v25) *
v26) * v18) - ((v25 * v18) * v23) + (((1 / 6 * v25) * v26) * v6) + (((1 / 2 * v25) * v23) * v6) + (((1 / 3 * v25) * v18) * v5) - (((1 / 6 * v25) * v6) * v5))) + (s11 *
((((2 / 3 * v25) * v26) * v6^(2)) - (((1 / 3 * v25) * v26) * v6) + (v25 * v23) - ((1 / 3 * v25) * v5))) + (s12 * ((((2 / 3 * v25) * v26) * v6^(2)) - ((v25 * v26) * v6) +
((1 / 3 * v25) * v26) + (v25 * v23) - ((1 / 3 * v25) * v5))) + (s13 * (((((1 / 3 * v25) * v26) * v23) * v6^(2)) - ((((1 / 2 * v25) * v26) * v23) * v6) + (((1 / 6 * v25)
* v26) * v23) - ((v25 * v17) * v23) + (((1 / 3 * v25) * v23) * v5))) + (s14 * (((((-1 / 3 * v25) * v26) * v24) * v6^(2)) + ((((1 / 2 * v25) * v26) * v24) * v6) + (((1 /
3 * v25) * v26) * v6^(2)) - (((1 / 6 * v25) * v26) * v24) + ((v25 * v17) * v24) - (((1 / 2 * v25) * v26) * v6) - (((1 / 3 * v25) * v24) * v5) + ((1 / 6 * v25) * v26) -
(v25 * v17) + ((1 / 3 * v25) * v5))) + (s15 * (((((1 / 3 * v25) * v26) * v23) * v6^(2)) - ((((1 / 6 * v25) * v26) * v23) * v6) - ((v25 * v17) * v23) + (((1 / 3 * v25) *
v23) * v5))) + (s16 * (((((-1 / 3 * v25) * v26) * v24) * v6^(2)) + ((((1 / 6 * v25) * v26) * v24) * v6) + ((v25 * v17) * v24) - (((1 / 3 * v25) * v24) * v5))) + (s17 *
((((1 / 3 * v25) * v26) * v6^(2)) - (((1 / 2 * v25) * v26) * v6) + ((1 / 6 * v25) * v26) - (v25 * v17) + ((1 / 3 * v25) * v5))) + (s18 * ((((1 / 3 * v25) * v26) *
v6^(2)) - (((1 / 6 * v25) * v26) * v6) - (v25 * v17) + ((1 / 3 * v25) * v5))) + (s19 * (-1)) + (sndg * (((v25 * v12) * v17) - ((v25 * v12) * v23) - ((v25 * v17) * v6) +
(((1 / 3 * v25) * v6) * v5) + ((1 / 2 * v25) * v23) - ((1 / 6 * v25) * v5)));
(* Contradiction! This proves the original statement. *)
(* The statement has a difficulty of degree 5. *)
```

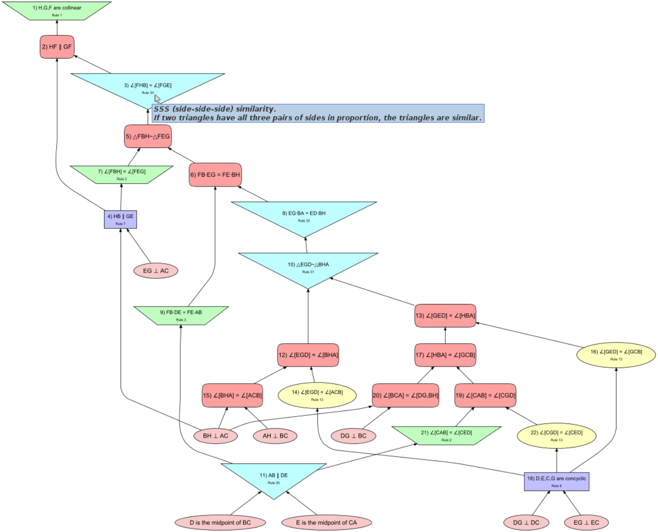
# Example 2: Euler's line

## Java Geometry Expert, Geometric Deductive Database Method



# Example 2: Euler’s line

## Java Geometry Expert, Geometric Deductive Database Method



# “Human” complexity

## Based on interviews with a limited number of students and teachers

- Interviewing 11 high school students (age 18) with focus on mathematics: Personal opinions on difficulty of 16 theorems, Szeged, Hungary, 2025 (Kerekes, K.)
- Master students in mathematics education from the Pontifical Catholic University of São Paulo, Brazil, 2025 (Abar, Vieira, R.)
- Master students in training teacher of secondary education (mathematics), Rey Juan Carlos University, Madrid, Spain, 2025 (T., R.)
- In-service mathematics teachers, workshop at V. Day of GeoGebra, Cuenca, Spain 2024 (R.)

# Comparison of difficulty of 14 classroom problems

Szeged (2025)

| Theorem | $C_a$ | $C_g$ | $C_h$ |
|---|---|---|---|
| Thales' ○ thm | 4 | 13 | 2.1 |
| Converse of Thales' ○ thm | 4 | 13 | 2.1 |
| Midline thm | 2 | 10 | |
| Intercept thm | 7 | 8 | 3.4 |
| Theorem of inscr. ∠ in a ○ | 7 | | 3.2 |
| ∠ of cyclic quadrilaterals | | | 3.2 |
| Rel. betw. inscr. ∠ and ∠ at the ⊙ | | | 3.2 |
| Tangent-secant thm | 6 | 16 | 3.8 |
| Wallace-Simson thm | 16 | 53 | 4.8 |
| Euler's line | 7 | 120 | 5.4 |
| 9-point ○ | 18 | 94 | 5 |
| Circumcenter of a △ | 2 | 25 | |
| Barycenter of a △ | 5 | 43 | |
| Apollonius' thm | 7 | | 3.2 |

# Comparison of difficulty of 10 classroom problems

## Cuenca, Madrid, São Paulo (2024–2025) — difficulty estimation also before proof attempt

1. Thales' Second Theorem: given two points $A$, $B$ and point $C$ on the circle with diameter $AB$, prove that $AC$ and $CB$ are perpendicular.
2. Prove that two lines parallel to a third one, are parallel to each other.
3. Given a line and a point outside it, prove that only one parallel line passes through that point.
4. Prove that if $D$ is the point of intersection of two altitudes of any triangle, then $D$ also belongs to the third altitude.
5. If $D$ is the point of intersection of two perpendicular bisectors on the sides of a triangle, then $D$ also belongs to the third perpendicular bisector.
6. The segments formed by the intersection of two parallel lines cut by two secant lines are proportional (according to Thales' Theorem).
7. Each cathetus (leg) is the geometric mean between the hypotenuse and its projection onto it.
8. The height relative to the hypotenuse is the geometric mean between the projections of the two catheti (legs) in a right triangle.
9. Prove that the line that passes through the midpoints of two sides of a triangle is parallel to the other side.
10. Prove that in a right triangle the square of the hypotenuse is equal to the sum of the squares of the catheti (legs) (Pythagoras).

# Correlation among different approaches of complexity

Pearson's $r$ ($-1 \leq r \leq 1$, inverse correlation – no correlation – positive correlation)

algebraic

# Correlation among different approaches of complexity

Pearson's $r$ ($-1 \leq r \leq 1$, inverse correlation – no correlation – positive correlation)

$$\text{algebraic} \overset{0.58}{\sim} \text{geometric}$$

# Correlation among different approaches of complexity

Pearson's $r$ ($-1 \le r \le 1$, inverse correlation – no correlation – positive correlation)

$$\text{algebraic} \overset{0.58}{\sim} \text{geometric} \overset{0.84}{\sim} \text{human}$$

# Correlation among different approaches of complexity

Pearson's $r$ ($-1 \leq r \leq 1$, inverse correlation – no correlation – positive correlation)

$$\text{algebraic} \overset{0.58}{\sim} \text{geometric} \overset{0.84}{\sim} \text{human} \overset{-0.07}{\sim} \text{AI}$$

# Correlation among different approaches of complexity

Pearson's $r$ ($-1 \leq r \leq 1$, inverse correlation – no correlation – positive correlation)

$$\text{algebraic} \overset{0.58}{\sim} \text{geometric} \overset{0.84}{\sim} \text{human} \overset{-0.07}{\sim} \text{AI}$$

What?! :-0

How is it possible that the difficulty corresponding to the length of AI's answer does not correlate with sophisticated machine algorithms or human intuition?

How is it possible that the difficulty corresponding to the length of AI's answer does not correlate with sophisticated machine algorithms or human intuition?

Possible answers:

- Wrong, incomplete or insufficient data.
  (Too small sample?)

## How is it possible that the difficulty corresponding to the length of AI's answer does not correlate with sophisticated machine algorithms or human intuition?

Possible answers:

- Wrong, incomplete or insufficient data.
  (Too small sample?)
- AI "thinks" it knows the solution and tends to explain it in a standard-length format...

## How is it possible that the difficulty corresponding to the length of AI's answer does not correlate with sophisticated machine algorithms or human intuition?

Possible answers:

- Wrong, incomplete or insufficient data.
  (Too small sample?)
- AI "thinks" it knows the solution and tends to explain it in a standard-length format...
- AI usually does not say: "I don't know".
  It tries its best — by attempting to turn
  "treasures from the trash heap" into a first-class meal...

## How is it possible that the difficulty corresponding to the length of AI's answer does not correlate with sophisticated machine algorithms or human intuition?

Possible answers:

- Wrong, incomplete or insufficient data.
  (Too small sample?)
- AI "thinks" it knows the solution and tends to explain it in a standard-length format...
- AI usually does not say: "I don't know".
  It tries its best — by attempting to turn
  "treasures from the trash heap" into a first-class meal...

Oversimplified answer:

- AI is not reliable

## How is it possible that the difficulty corresponding to the length of AI's answer does not correlate with sophisticated machine algorithms or human intuition?

Possible answers:

- Wrong, incomplete or insufficient data.
  (Too small sample?)
- AI "thinks" it knows the solution and tends to explain it in a standard-length format...
- AI usually does not say: "I don't know".
  It tries its best — by attempting to turn
  "treasures from the trash heap" into a first-class meal...

Oversimplified answer:

- AI is not reliable (at least in this manner)

## How is it possible that the difficulty corresponding to the length of AI’s answer does not correlate with sophisticated machine algorithms or human intuition?

Possible answers:

- Wrong, incomplete or insufficient data.
  (Too small sample?)
- AI “thinks” it knows the solution and tends to explain it in a standard-length format. . .
- AI usually does not say: “I don’t know”.
  It tries its best — by attempting to turn
  “treasures from the trash heap” into a first-class meal. . .

Oversimplified answer:

- AI is not reliable (at least in this manner) ¯_(ツ)_/¯

# Focused features of the mentioned software tools

*ChatGPT*, second approach with *more accurate input*

- User input:
  - hypotheses + thesis in `DescribeStatement` format
- Output:
  - A textual proof (more reliable)
  - Complexity grade measured
    via character length of the text output

# Correlation among different approaches of complexity

## Second attempt with more rigorous input, via `DescribeStatement`

$$\text{algebraic} \overset{0.41}{\sim} \text{AI}$$

$$\text{geometric} \overset{0.37}{\sim} \text{AI}$$

$$\text{human} \overset{0.50}{\sim} \text{AI}$$

# Correlation among different approaches of complexity

Second attempt with more rigorous input, via `DescribeStatement`

$$\text{algebraic} \overset{0.41}{\sim} \text{AI}$$

$$\text{geometric} \overset{0.37}{\sim} \text{AI}$$

$$\text{human} \overset{0.50}{\sim} \text{AI}$$

→ moderate correlation after refinement

# What did we learn from this project?

- Using AI is nice, but human checking of its answer is a must.
- Verification and refinement seems to require human assistance. Automating the automation seems to require human creativity.
- Forcing the human user to enter more detailed input can be crucial to obtain more reliable output.
- Forcing AI to use **strong formalized output** may be the next step to double-check the proofs it produces. (Double-checking should be performed with *another system*.)
- A first-round check of difficulty of a geometry problem via AI can be a good approximation, but a second check with a white box algorithm still seems necessary.
- It is mandatory to continue developing simple and verifiable methods to make it possible to double-check the results obtained via AI.

“People constantly try to offload their thinking,
yet they never fully succeed.
It seems we’re simply designed to keep thinking.”

“People constantly try to offload their thinking,
yet they never fully succeed.
It seems we’re simply designed to keep thinking.”
— A Human But Forgotten Author
(AI does not find the quotation source, sorry)

# Acknowledgments

Other people who contributed to this presentation include:

- Celina Abar (Pontificia Univ. Católica de São Paulo, Brazil)
- András Kerekes (M. Radnóti High School, Szeged, Hungary)
- Belén Ariño-Morera (Univ. Rey Juan Carlos, Madrid)
- Angélica Martínez-Zarzuelo (Univ. Complutense, Madrid)
- Álvaro Nolla (Univ. Autónoma de Madrid)

The authors gratefully acknowledge the support of the project "Augmented intelligence in mathematics education through modeling, automatic reasoning and artificial intelligence" (IAxEM-CM, `https://iaxem.es`), PHS-2024/PH-HUM-383, granted by the Comunidad de Madrid.

# References I


- Abar, C. A. P. , Almeida, M. V., Recio, T.: "Estimating the Complexity Levels of a Geometric Statement". Boletín de la Sociedad Puig Adam, no. 121, pp. 67–88, (2026).
- Ariño-Morera, B.: GeoGebra Discovery at EGMO 2022. Revista do Instituto GeoGebra Internacional de São Paulo, ISSN-e 2237-9657, Vol. 11, no. 2, pp. 5–16, (2022).
- Ariño-Morera, B., Kovács, Z., Recio, T., Tolmos, P.: "Solving with GeoGebra Discovery an Austrian mathematics olympiad problem: Lessons learned". In: Quaresma, P., Kovács, Z. (eds.) Proceedings 14th International Conference on Automated Deduction in Geometry, Belgrade, Serbia, 20-22th September 2023. Electronic Proceedings in Theoretical Computer Science, vol. 398, pp. 101–109. Open Publishing Association, (2024).
- Bartha, S., Kerekes, A., Kovács, Z., Recio, T.: "Automated analysis of the difficulty of secondary school geometry theorems". Ann. Math. Artif. Intell., 93, pp. 1011–1033, (2025).
- Botana, F., Recio, T.: "Geometric Loci and ChatGPT: Caveat Emptor!". Computation, 12(2), 30, (2024).
- Botana, F., Recio, T., Vélez, M. P.: "On Using GeoGebra and ChatGPT for Geometric Discovery". Computers, 13(8), 187, (2024).

# References II


- Kovács, Z., Parisse, B., Recio, T., Vélez, M.P., and Jonathan H. Yu: “The ShowProof command in GeoGebra Discovery: Towards the automated ranking of elementary geometry theorems.” ACM Communications in Computer Algebra, Vol. 58, No. 2, Issue 228, June 2024. pp. 27–30. (2024).
- Kovács, Z., Recio, T., Vélez, M.P.: “Automated reasoning tools with GeoGebra: What are they? What are they good for?” In: Richard, P.R., Vélez, M.P., Van Vaerenbergh, S. (eds.) Mathematics education in the age of artificial intelligence: how artificial intelligence can serve mathematical human learning, pp. 23–44. Springer, (2022).
- Quaresma, P., Graziani, P.: “Measuring the Readability of Geometric Proofs: The Area Method Case.” J. Autom. Reasoning 67, 5., (2023).
- Quaresma, P., Graziani, P., Nicoletti, S.M.: “Considerations on Approaches and Metrics in Automated Theorem Generation/Finding in Geometry.” In: Quaresma, P. and Kovács, Z. (Eds). Proceedings 14th International Conference on Automated Deduction in Geometry. Electron. Proc. Theor. Comput. Sci. 398, Open Publishing Association, 85–100. https://doi.org/10.4204/EPTCS.398.12 (2024).
- Wos, L.: Automated Reasoning: 33 Basic Research Problems. Prentice-Hall, (1988).