

# Hungarian Riddles Benchmark

## A No-Code Evaluation Project

*Developed by Károly Boczka*

*Multilingual AI Evaluator & Data Analyst — October 2025*

---

### Project Overview

This project presents a Hungarian benchmark evaluation designed to practice and demonstrate no-code model evaluation skills. The dataset contains 100 metaphorical, trivia-style riddles in Hungarian, covering topics such as culture, history, sports, geography, daily life, politics, arts, and notable Hungarian figures. These riddles are intentionally tricky — built to test factual accuracy, reasoning clarity, and language quality.

Two native Hungarian participants helped to establish **human reference baselines**:

Baseline A – university-educated male, 50+ years, 95/100 correct

Baseline B – 18-year-old male, 50/100 correct

These baselines served as reference points for LLM comparison.

### Evaluation Setup

The same 100 riddles were presented to six large language models (LLMs). Each model was instructed to:

- Respond in Hungarian
  - Provide a concise justification ( $\leq 2$  sentences)
  - Avoid speculation or multiple alternatives
  - Use web search when possible
  - Provide one final answer directly, without clarifying questions
- 

### Evaluation Rubric (0–5 scale per category)

#### **Factual Accuracy** (0 / 1 / 3 / 5)

0 = no answer / irrelevant | 1 = on-topic but wrong | 3 = partially correct | 5 = fully correct

#### **Justification** (0–5)

0 = hallucination | 1 = no justification | 3 = partial reasoning | 5 = concise, logical, adequate

#### **Grammar & Style** (0–5)

Grammar (0–3): 0 = unintelligible / not Hungarian; 3 = correct and fluent

Format (0–1): 1 =  $\leq 2$  sentences, concise

Style (0–1): 1 = clear and culturally appropriate

Auto-0 Rule: If not in Hungarian or no justification  $\rightarrow$  total = 0

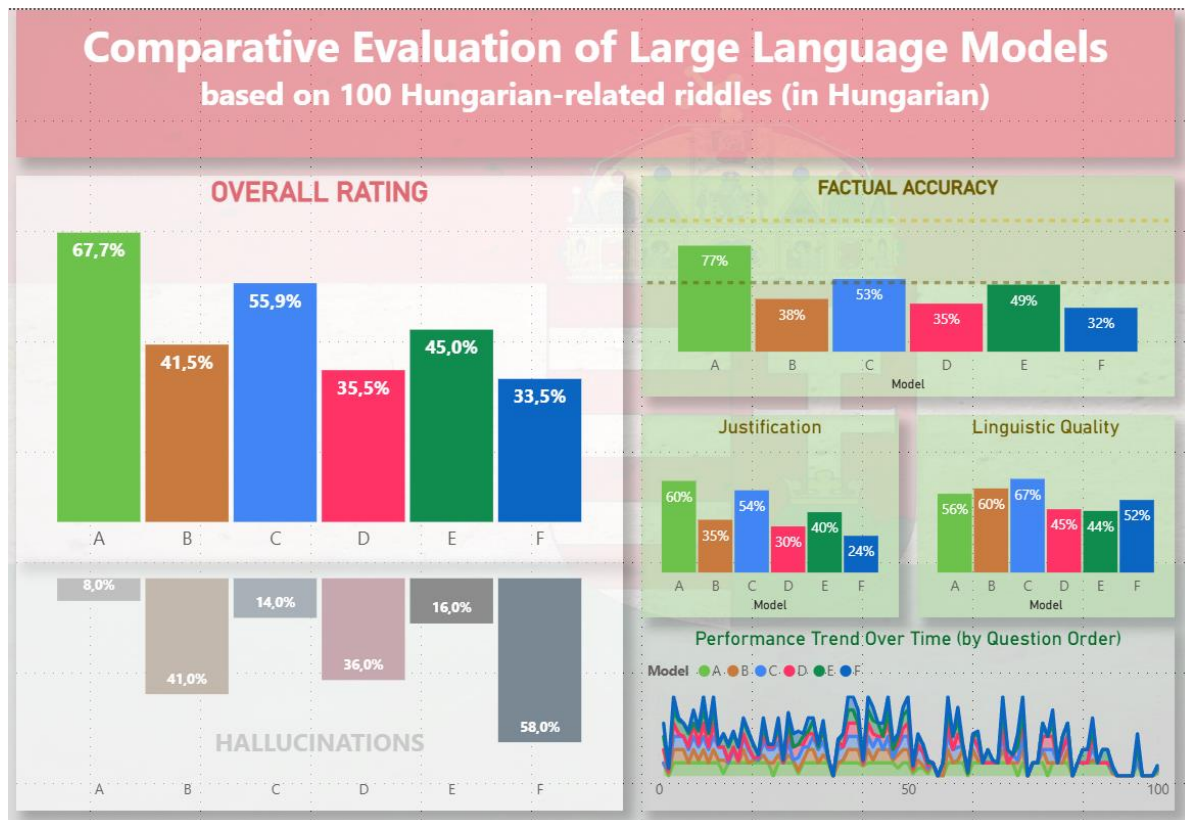
Scoring followed the rubric above and was analyzed using Power BI dashboards.

---

## Model Selection (September 2025)

GPT-5 Deep Search (A), Claude Sonnet 4 (B), Gemini 2.5 Flash (C), Grok (D), GPT-5 (E)  
DeepSeek R1 (F)

This setup allows comparison between standard GPT reasoning and search-augmented GPT, while representing diverse architectures and response styles.



## Key Findings

On **factual accuracy**, no model surpassed the higher human baseline (95%). Only Model A (77%) exceeded the lower baseline (50%), while Models C (53%) and E (49%) performed comparably. Models B (38%) and D (32%) fell below the lower human mark.

No single model dominated **all categories**. **Model A led consistently, dominated** in factual accuracy and consistency. **Models C and E performed mid-tier**, balanced reasoning with fluency. **Model B's language fluency was exceptional but often inaccurate** — confirming a trade-off between eloquence and factual precision. **Models D and F were weaker overall**, Model F underperformed due to a high hallucination rate standing out as a clear limitation.

Performance declined slightly over the 100 riddles, likely due to session fatigue or increased difficulty. The Power BI dashboard trendline visualizes this gradual drop.

Culturally loaded riddles produced elevated hallucination rates: even the strongest model hallucinated in 8% of responses, while the weakest exceeded 50%.

## Model Performance Summary

Model	Overall	Factual	Justification	Language	Hallucination
A	67.7%	77%	60%	56%	8%
B	41.5%	38%	40%	66%	41%
C	55.9%	53%	45%	67%	14%
D	35.5%	35%	35%	45%	36%
E	45.0%	49%	30%	44%	16%
F	33.5%	32%	24%	52%	58%

---

## Limitations

This was a single-rater, single-session benchmark, designed as a pilot rather than a scientific study. Later runs will test multi-prompt sessions, larger sample sizes, and HU-EN-HR multilingual extensions. While the setup revealed session drift (performance decline over time), it also means later scores may underrepresent each model's full potential.

---

## Data Access

All raw responses (600 items), answer keys, and evaluation tables are available in my GitHub repository.

---

## Framing Disclaimer

This is not a definitive benchmark or scientific study. It is a personal, independent pilot project created solely to demonstrate evaluator methodology — structured task design, consistent rubric use, and clear result communication.

Model names are disclosed once for transparency but anonymized (A–F) throughout to focus on evaluation quality rather than product comparison.


---

## Conclusion

This benchmark confirms that factual and cultural nuance often diverge in smaller languages. Human-in-the-loop evaluation remains vital for smaller languages like Hungarian and culturally specific reasoning tasks. While top models approach fluency, human oversight continues to ensure interpretive depth and factual grounding.

---

## Copyright & Credits

© 2025 Károly Boczka — mention me in goodwill, or buy me a virtual beer 

All materials are freely reusable with credit.

# APPENDIX

## Master Prompt

You are asked to solve 100 short Hungarian riddles or quiz-style prompts, each reflecting Hungarian culture, history, everyday life, sports, politics, literature, and other locally relevant domains.

## Answering Rules

**Language:** Always answer in Hungarian.

**Length:** Each answer must be maximum 2 sentences.

**Justification:** Include a 1–2 line factual justification (clear, concise, no verbosity).

**Web Search (Mandatory):** For every riddle, you must perform a web search before answering, regardless of category or confidence. Do not rely only on internal knowledge.

**Fallback Rule:** If no reliable result is found after searching, answer with: *“Nem tudom.”*

**Process Discipline:** Attempt one best-effort solution directly; no clarifying questions.

**Evaluation Context:** These riddles are intentionally tricky. They are designed to test factual accuracy, reasoning clarity, and language quality.

## THE 100 RIDDLES

### Hungarikums

Eme vöröskeresztes gömb nedve nem édes a hasadnak.

Meséld el a viccet, amiben keresztnevek és városnevek keverednek, mert egy kisiú nem tudja, hogy hívják és miért küldték a boltba.

Ezt a téstát a kémény belsejében készítik és nem hangszerre hasonlít.

Édes, úri szobákban lenne a helye szép csillogó ruhájában, de inkább fákon lóg.

Nagy természetű, növényevő hím testnedvét látják eme hevesi nedűben.

Még az uszodáját is jellegzetes fűszernövényéről formázta ez a városka.

Kócos kis jószág, azt sem tudni melyik az eleje, melyik a hátulja, mégis jól mutatja az irányt nála jóval nagyobbaknak.

Hajlított asztalon focilabdával lábtengőzni.

Egyes magyar városokban - eltérő módokon - folyami állatokból folyékony ételeket készítenek.

Rövid ideig élő állatok látványos násztánca, mégis növényeknek hívjuk őket.

### Kids' World

Ki az a gyorslábú kis vörös, akinek az útjából a vadászok jobban teszik, ha kitérnek.

Egy rendőr, akinek egy kutya volt a szirénája.

Egy zöld szörny folyton arról énekel, hogy virág szeretne lenni.

Kedves, lelkesen szerel, de rendetlen cselekedete nem nyer, kellemetlen helyzetekbe kevered, cégére rendre leverett.

Messzi földről származó őskori alakok, akik remek rímekben szólalnak meg magyarul.

Ha akarom, kifordított bundakesztyű, szobafestő pemzli vagy papucs orrán pamutbojt.

Papa, mama, gyerekek, csupa szív, szeretet, egy se nyafog, kesereg, de kik ők?

Egy zöld kisállat, akinek mértani alakzatokból áll a füle, azzal repül.

Szeretem egy bizonyos édesség minden fajtáját, emiatt már repülni sem tudok.

Egy vicces, háromszemű robot vagyok, és kisgyerekek társaságában keveredek galaktikus kalandokba.

## History

Mi volt a hazát szerző magyarok jellegzetes harci taktikája, mit tudtak négy lábúak hátáról nagyon jól csinálni?

A monda, amiben egy magyar megölte ellenségét egy hangszerrel.

A magyar hős egy-egy máshitút a két hóna alá csapott, egyet a foga közé szorított és úgy járta.

Hogy hívták születésekor azt híres ősrünket, akit egy napon ünnepelünk az új kenyérrel?

Egy bátor magyar, aki fejest ugrott a hazáért, miközben ölelt egy idegen jelképet. Nem is volt magyar, sőt, nem is létezett.

Ha megkondul a nap derekán, mind tudjuk az okát, valaki győzött valahol.

Korábban bolgár, tejszínű erődítmény volt, ma már más nyelvet beszélnek itt.

Sötét erők támogatták, de mikor meghalt, a hazugság ellenfeleként siratták.

Szerencsétlen szám, főleg ha nyakkendő helyett más van a nyakad körül.

Bátor suhancok egy ellenségről elnevezett "tüzes itallal" támadtak idegenből érkező "nagyobb testvéreikre".

## Sports

Gyorsléptű katona, egész magas rangig vitte. Mégis az egész ország tegezte és imádta, pedig végül nem is itt kergette a pettyest.

Megtelt a betonteknő, mikor a rendőrök és a katonák népszerű labdazsonglőrei csaptak össze.

Ugyan kettőnél többet kaptunk, de a dupláját adtuk cserébe, gombócból is sok, azóta is ünnepeljük.

A magyar futball Trianonja és Mohácsa, fél tucattal a zsákban.

Ha a kolbászosok és a fényképesek egymásnak esnek, nem a rúgásoké a főszerep.

Nyíl és Törő egymás ellen, de melyik klubokról van szó?

Döme a világ egyik akkori legjobbja, Germániában, Itáliában és Hellászban is sztár volt.

Háromszor egymás után a dobogó tetején az ötkarikás játékokon, ez csak egy csapatunkak sikerült.

A földgolyó másik oldalán egy úszómedencében keményen álltak ki megtámadott hazájukért a magyar legények.

Salzburgból Lipcsébe, onnan meg a ködös Albionba terelgettem a labdát.

## Geography & Places

Hogyan lehet eljutni az Osztjapenkótól a Felszabra?

A reformátusoknak olyan ez a város, mint hitbéli társaiknak Csizmaország központja.

Itt könyeret möggyel ösznek és korábban meggyűlt a bajuk a vízzel.

Melyik az a híres budapesti építmény, amit olyan állat őriz, amelyiknek hiányzik a nyelve

Korábban egy bajszos idegen nevét viselte, ma már egy vízfolyását ez a település

Ez a dunántúli hely visszamondja neked, amit kiabálsz, de vigyázz, a szerzetesek is hallják.

Hogy jutok legkönnyebben a Hősökről a Moszkvára?

Egy hely, ahol az első magyar belépett a Kárpát-medencébe és egy műalkotás, ami erről készült.

Valahol lóháton állva kancsikát csattogtatnak, így igazítják a nagytestű kérődzőket útba.

Bár korábban három tenger mosta partjaink, mára nekünk ez maradt, de nagyon szeretjük.

## Everyday Life

Mennyire volt jókedvű a piros nyakkendősrác a nyolcvanas években?

Egy kártyajáték, amelynek másik neve egy kétjegyű szám, amelyben mindkét jegy ugyanaz.  
Egy hosszú lépéssel vagy egy távoli ugrással kerülünk közelebb a mámoros célhoz?  
Mi volt a nyolcvanas években a magyar fizetőeszköz legkisebb címletének hátulján?  
Egy állat, akiről bulvármagazint neveztek el a rendszerváltás után.  
Harminc éve csak egy ötvenest értem, de ma már egy ötszázast.  
Kicsoda Stüsü vadász és Tök filkó, mi a közös bennük?  
Találmányommal játszanak kicsik és nagyok, pedig geometriai formákat kell csak forgatni.  
Ezzel az itallal nem ütjük össze poharainkat, bár egyesek szerint már lejárt a tilalmi időszak.  
Imádtuk ezt a játékot, dobtunk a kockával és lett minden a lakásba, amire vágytunk.

## **Politics, public life**

Nemcsak a mezőgazdaságból élőknek, de egy zöld-fehér sportklubnak is főnöke lett.  
A déli féltekére akart egy ott is honos gyümölcsöt exportálni ez a politikus.  
A magyar tenger partján mondott beszédet egy politikus, nem kellett volna.  
Elismerte, hogy néhány napszakban nem mondtak igazat, és nem ő volt az első ilyen.  
Egy tornászlány, aki később sporteseményeket közvetített, de a magyarok egy nagy részének más jut róla eszébe.  
Kedvenc szeszesitaláról elnevezett hírhedt bűnöző.  
Női keresztnéve ellenére egy kegyetlen külföldi bérgyilkos, kiírtott egy ismert alföldi családot.  
Egy idősebb úrnak szóló kedves gyermeki köszöntést is jelenthetne ennek a hazánkban is rettegett ukrán bűnözőnek a neve.  
Utasokat szállítók maguk akadályozták a forgalmat a székesfőváros útjain.  
Egy juhász nem csak focipályát épített, de még az utolsó vacsorát is eladta, el is zárták.

## **Literature / Arts**

Viccek agyafúrt gyerekszereplője, '90 előtt így általában így hívták, utána meg úgy.  
Egy hosszú faágat nyújt ki karjával, úgy mutatja a pökhendi idegen lovasnak merre van az arra.  
Kamaszok védik a játszóterüket a színes felsőruhás nagyfiúk ellen, közben egyiküket árulással vádolják és szegény meg is hal.  
Egy kotlós nagyon jól érzi magát a házban, a Teremtő jól rendezte el a sorsát, de egy kutyanak el kell magyarázni, hogy nem szabad őt bántania.  
Valakinek egy kés van a hátában, de nem adja vissza a jogos tulajdonosának. Egy slamos öreg meg csak kavar a háttérben, miközben csattognak a pofonok.  
A főhős magas tudományos kitüntetést nyer kártyán. Később titkárául szegődik egy fura alak, akit kitüntetnek helyette egy katonai szervezetben is. Egy igen értékes autó is központi szerepet kap.  
Nem csillagokról szól. Két kisgyereket messzire hurcolnak otthonról, de később egy óriási birodalom központjából is megszöknek, hősieken küzdenek ellene a vár fokán.  
Egy furcsa szívű ember gyermekeinek romantikus kalandjai.  
Egy kemény, de tisztességes öreghez kerülnek a kamaszfiúk, akik megemberesedve tértek haza a balatoni nyári vakációból.  
Fél tucat és egy napja csak anyukám jár a fejemben. Megy felfelé egy lavórt cipelve a ház felső szintjére, ősz haja lobban.

## **Movies**

Egy bátor nő folyton megveri a rosszakat és közben sikítozik, meg kismotorral száguldozik.  
Ugyanott lakik egy kopasz mentős, egy ladás taxis, egy nyomdász fiú és még sokan mások.  
Mikiék elvileg nem ellenségek, sőt rokonok is, mégis keresztbe tettek egymásnak több mint 20 éven át. Egy ország nézte, senki sem ismerte be, hogy nézi.  
Ruhaakasztókkal kereskedő a lóspart szerelmese, unokaöccsét felügyeli, mennek mindenfelé.

Hősünk csak arra szeretne vigyázni, hogy a folyó ne öntsön ki, de mindenféle politikai játszmákba rángatják, még börtönbe is kerül.

Anyósával élő vasúti felügyelő alkalmaz egy tehenész fiút, és ételdobálásba fullad a sok kalamajka.

Igazából gyerekeknek készült, de felnőttek is imádják a cincogók és nyávogók közötti csatározásokat, meg a vérszívók zenekarát.

Az egyik dombság vezetője lóra pattan, hogy sorozatosan borsot törjön az elnyomó sógorok orra alá.

Nagykamaszok kalandjai, egy ikonikus fővárosi helyszínről (amit egy másik fővárosról neveztek el) indul mindig a tuti buli.

Egy államférfi megbotlik, mert nem veti meg a másik nemet és ez rengeteg bajba sodorja őt is, környezetét is, egy pesti belvárosi szálloda kulisszái között.

## **Music**

Egy szép napon, biztos vagyok benne, hogy elmegyek a szülővárosomból, és minden ott marad, amit tőle kaphattam.

Csoroghat, futhat az a kis csermely. Távozni tervezek, de még nem tudom merre, távozok eddigi otthonomból, még azoktól sem köszönök el, akik a legközelebb álltak hozzám.

Ha azt hiszed, hogy létedet jutalmul kaptad, tekints őseidre, miként élnek - egy túlvilági utazó.

Ha legyőzöd ellenséged, aki a rokonod, hatalmas nagy lesz az ország, határai minden irányban kitolódnak, történelmi nagyjai nem hős vesztesek, hanem dicső győztesek lesznek.

Fiatal csajszi a hangszer mellett, sápadtabb, mint egy másik hangszer.

Nagymacskák, nagy és kistestű majmok, kávébarna hölgyek, egy földrész.

Kezdődjön a muri, hulljon darabokra az épület, töltsön meg minket a jelenlegi hevület.

Több száz esztendő's játék már ez, csak rájönnek mire lehet használni, távol-keleti országokból érkezett. Ne akarj olyan felnőtt férfinak tűnni, inkább jöjj és próbáljuk ki.

Lassan sétál a fegyveres erők tagja, fogalma sincsen neki sem, hogy hova megy. Ajkai között egy fura hangszer.

Fényes égitestek, meséljétek el, hol találom azt, aki közel áll a szívemhez.