The CLEAN PAPER

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

የቋንቋ ሞዴሎች ለምን ቅዠት ያደርጋሉ — የምንመዝናቸው መንገድስ ለምን እንዲቀጥል ያደርጋል?

ቅዠት የምንለው በእርግጠኝነት የሚነገር ሐሰት ምስጢራዊ ብልሽት አይደለም፤ አንዳንዱ የስልጠና ስታቲስቲካዊ ውጤት ነው፣ እና ዋና benchmarks ከታማኝ "አላውቅም" ይልቅ እርግጠኛ ግምትን ስለሚሸልሙ ይቀጥላል። በአራት ግንባር ቀደም ሞዴሎች የተደረገ የጉዳይ ጥናት፣ የውጤት ሕጎቹን በprompt ውስጥ መግለጽ ("open rubrics") ይህን ማበረታቻ እንደሚገለብጥ ያሳያል።

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · AI-assisted, human-reviewed

Paper: *Evaluating large language models for accuracy incentivizes hallucinations*, Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang, Nature 653, 1047–1050 (2026) · DOI: 10.1038/s41586-026-10549-w

ሞዴሎች ለምን ይገምታሉ፣ እኛስ ለምን እንዲገምቱ አስተማርናቸው?

ትልቅ የቋንቋ ሞዴልን የማያውቀውን ሰው የትውልድ ቀን ጠይቀው፤ “ማርች 7” ብሎ ከካርድ እያነበበ እንደሚናገር ሰው በጸና እርግጠኝነት ሊመልስ ይችላል — እና በተከታታይ ሦስት ጊዜ፣ በሦስት የተለያዩ ቀናት፣ ሊሳሳት ይችላል። ደራሲዎቹ በትክክል ይህን ዓይነት ምሳሌ ይሰጣሉ፤ ግንባር ቀደም ሞዴሎች ቀላል የእውነታ ጥያቄ — የአንድ ሰው የትውልድ ቀን፣ ወይም እምብዛም የማይታወቅ ምሕፃረ ቃል ምን ማለት እንደሆነ — ሲጠየቁ፣ እያንዳንዳቸው በእርግጠኝነት የተለየ መልስ ፈጠሩ፣ አንዳቸውም ትክክል አልነበረም። ኢንዱስትሪው ይህን ቅዠት ይለዋል፣ ይህም የአመለካከት ችግር እንደሆነ ያስመስለዋል። የጽሑፉ የመጀመሪያ እርምጃ ምስጢሩን ማስወገድ ነው።

ሞዴል እንዴት እንደሚገነባ እንጀምር። በመጀመሪያውና በትልቁ የስልጠና ደረጃ፣ እጅግ ብዙ ጽሑፍ በማንበብ ቅልጥፍና ያለው ቋንቋ ምን እንደሚመስል ይማራል። አሁን ከጀርባው ምንም ሥርዓት የሌለውን እውነታ — የአንድ የተወሰነ ሰው የትውልድ ቀን — እንውሰድ። ቀኑ በስልጠና ጽሑፍ አንድ ጊዜ ብቻ ከታየ፣ ወይም በፍጹም ካልታየ፣ ሥርዓት የሚማረው ሞዴል የሚይዘው ነገር የለም፤ ከሞዴሉ አንጻር መልሱ የዘፈቀደ ነው። ደራሲዎቹ ይህን በአሮጌ ሐሳብ (የAlan Turing፣ ከሌላ ችግር) በመጠቀም ያብራራሉ፤ ከአምስት የትውልድ ቀናት አንዱ በመረጃው ውስጥ አንድ ጊዜ ብቻ ከታየ፣ ሞዴሉ ቢያንስ ከአምስት አንዱን እንዲሳሳት ይጠበቃል — ስለተበላሸ አይደለም፣ ከመጀመሪያው የሚማረው ነገር ስላልነበረ ነው። (በዚህ አመክንዮ ሞዴሎች የአገር ዋና ከተማን እምብዛም አይሳሳቱም፤ እነዚህ ብዙ ጊዜ ይታያሉ።) እውነተኛ መግለጫን ከሚመስል ሐሰት መለየት ራሱ ከባድ ችግር እንደሆነ፣ እውነተኛ መግለጫዎችን ብቻ ማመንጨትም ቢያንስ እንደዚያ ከባድ እንደሆነ በጥንቃቄ ይከራከራሉ። ዝቅተኛ የስህተት ገደብ በሥርዓቱ ውስጥ ተገንብቷል።

ከሥሩ ያለው ሐሳብ፦ እስካሁን ያላዩትን እንዴት መቁጠር ይቻላል?

ይህ በእውነት ብልህ በሆነ ሐሳብ ላይ ይመሠረታል — ከቋንቋ ሞዴሎች የቆየ፣ በትክክል ሊታወቅም የሚገባው። **ባለቀለም ኳሶች ባሉበት ከረጢት ይጀምሩ።** ስንት ቀለሞች እንዳሉት አያውቁም። 100 ኳሶችን አንድ በአንድ አውጥተው ይቆጥራሉ፦ ቀይ 40፣ ሰማያዊ 25፣ አረንጓዴ 15፣ ቢጫ 5፣ ሐምራዊ 3፣ ብርቱካናማ 2 — ከዚያም **እያንዳንዳቸው አንድ ጊዜ ብቻ የታዩ አሥር የተለያዩ ቀለሞች።**

አሁን Turing በሌላ ችግር ላይ ያጋጠመውን ጥያቄ እንውሰድ፦ *ቀጣዩ ኳስ በፍጹም ያላዩት ቀለም የመሆን ዕድል ምን ያህል ነው? ያላወጡትን ነገር መቁጠር አይችሉም — ግን አንድ ጊዜ ብቻ ያዩአቸውን “singletons” መቁጠር ይችላሉ።* Good–Turing estimation የሚባለው ዘዴ፣ ከወጡት ኳሶች ውስጥ singleton የሆኑትን ድርሻ በመጠቀም ባላዩአቸው ቀለሞች ውስጥ የተደበቀውን ዕድል ይገምታል። ከመቶው ኳሶች አሥሩ አንድ ጊዜ ብቻ የታዩ ቀለሞች ነበሩ፣ ስለዚህ ቀጣዩ ኳስ ፍጹም አዲስ ቀለም የመሆን ዕድል 10 / 100 = 10% ያህል ነው።

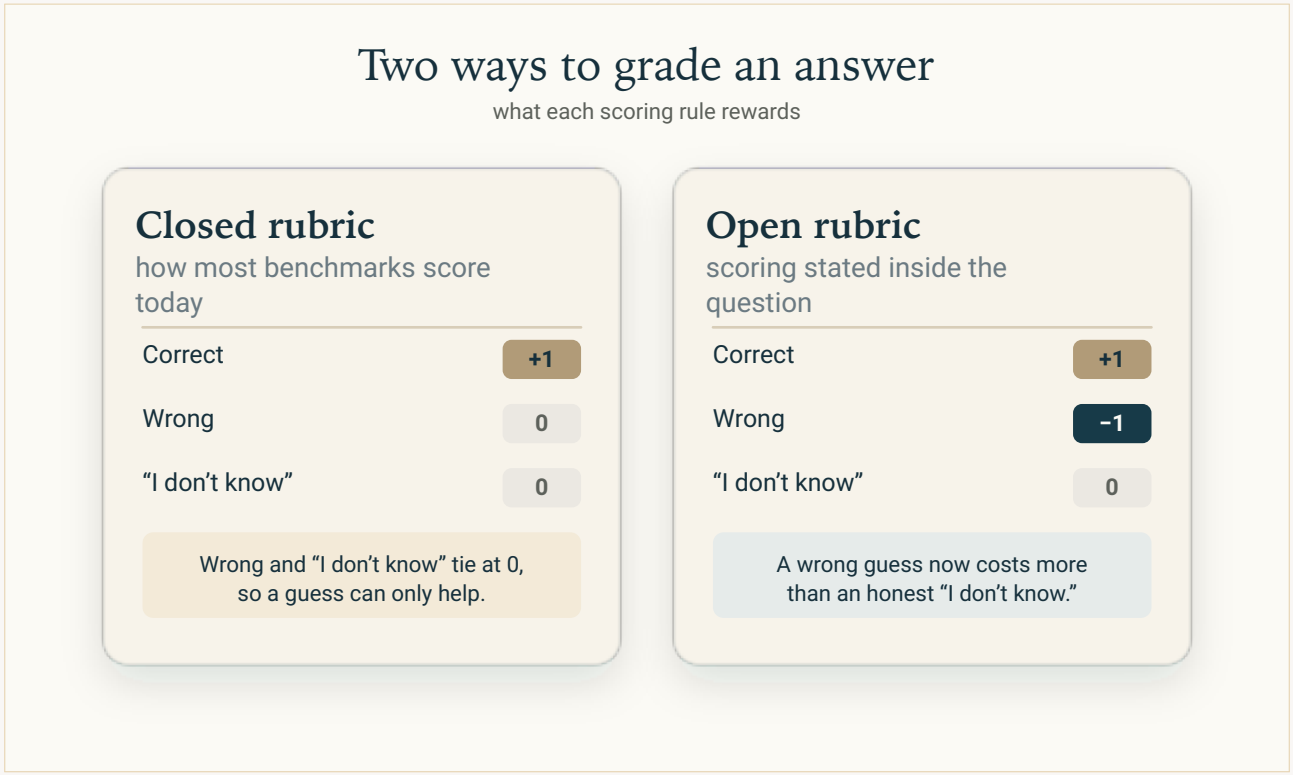
እነዚህ አንድ ጊዜ የታዩ ቀለሞች ስህተቶች አይደሉም። የራስዎን አለማወቅ ይለካሉ፤ ብዙ ቀለሞች አንድ ጊዜ ብቻ ሲታዩ፣ ናሙናው እስካሁን ያላወጡአቸው ብዙ ነገሮች እንዳሉ እየነገረዎት ነው።

አሁን ቀለሞቹን በትውልድ ቀናት፣ ከረጢቱንም በሞዴሉ የስልጠና ጽሑፍ ይተኩ። ሞዴሉ ካያቸው ቀናት ውስጥ **ከአምስት አንዱ አንድ ጊዜ ብቻ ይታያል** እንበል። ዘዴው ያው ነው፤ ከዕድሉ አምስተኛው ያህል ሞዴሉ በተግባር ባላያቸው ቀናት ውስጥ ነው — የትውልድ ቀንም የሚደገፍበት ሥርዓት የለውም (የሰውን የትውልድ ቀን ማስላት አይቻልም)። ስለዚህ አንድ ጊዜ የታየ ወይም በፍጹም ያልታየ ቀን ሞዴሉ ሊመዝነው የማይችለው ሳንቲም ነው፣ እና **ከአምስት አንዱን** ያህል ይሳሳታል። ምንም ያህል ብልህነት ይህን አያስተካክለውም፤ የሚማረው ነገር አልነበረም።

ክርክሩ በአጭሩ ይህ ነው፦ singleton መጠን *ከዚህ መረጃ* ሊማር የማይችለውን የዓለም ክፍል ይለካል፣ ይህም ለስህተቶች **ዝቅተኛ ገደብ** ይሆናል። ሞዴል ዋና ከተማን እምብዛም የማይሳሳተውም ለዚህ ነው — *Paris* ብዙ ጊዜ ይታያል፣ singleton መጠኑ ወደ ዜሮ ይቀርባል፣ ስለዚህ ብዙ የሚማር አለ።

ይህ ቅዠት ከየት እንደሚመጣ ያብራራል። ለምን እንደሚቀጥል ግን አያብራራም — ሞዴሎች ጠቃሚና ታማኝ እንዲሆኑ

ከተደረገው ተጨማሪ ስልጠና በኋላ እንኳ፣ ለምን ጥርጣሬያቸውን ከመቀበል ይልቅ እንደሚያስመስሉ። የጽሑፉ ምሳሌ ምኞት እስኪያሳጣ ድረስ ተስማሚ ነው። በፈተና ላይ መልስ የማያውቅ ተማሪን ያስቡ። ባዶ ቦታ ዜሮ ካገኘ፣ ግምት ግን አንድ ሊያገኝ ከቻለ፣ ውጤትን የሚያሳድገው ግምት መስጠት ነው — በእርግጠኝነት፣ በተወሰነ መልስ፣ “እርግጠኛ አይደለሁም” ሳይል። ተማሪዎች ይህን ይማራሉ። እኛ በዚያው መንገድ ስለምንመዝናቸው ሞዴሎችም ይማራሉ። ደራሲዎቹ መስኩ በሚወዳደርባቸው benchmarks እና ሞዴሎች እንዲወጡባቸው በሚያስተካክላቸው leaderboards ውስጥ ተመለከቱ፣ እና ከሞላ ጎደል ሁሉም “አላውቅም” ለሚለው ልክ እንደተሳሳተ መልስ ዜሮ እንደሚሰጡ አገኙ። በዚህ ሕግ፣ ሁልጊዜ የሚገምት ሞዴል ጥርጣሬውን በታማኝነት ከሚገልጽ ተመሳሳይ ሞዴል ይበልጣል። በቃል ትርጉም የሚቀርብ ያህል፣ የምንሰጠው ውጤት ወደዚህ ባህሪ ይገፋቸዋል።



Benchmarks ግምት መስጠትን ምክንያታዊ ሊያደርጉ ይችላሉ። ብዙ benchmarks በሚጠቀሙበት አሰጣጥ (ግራ)፣ የተሳሳተ መልስ እና ታማኝ “አላውቅም” ሁለቱም ዜሮ ያገኛሉ — ስለዚህ መገመት ሊረዳ ብቻ ይችላል። ሕጎቹን በጥያቄው ውስጥ ግልጽ አድርጉ፣ ለስህተትም ቅጣት ይኑር (ቀኝ)፤ እርግጠኛ ሳይሆኑ መቆጠብ የተሻለ ምርጫ ሊሆን ይችላል። ይህ ፈተናው የሚሸልመውን ይቀይራል፤ በራሱ ግን ቅዠትን አይፈታም።

የተለመደውን ርዕስ ስለሚቃወም ሊያዝ የሚገባው ክፍል ይህ ነው። ቅዠት ብዙ ጊዜ የቴክኖሎጂው የማይቀርና ምስጢራዊ ወሰን ተደርጎ ይቀርባል። ጽሑፉ ሁለቱንም ይቃወማል። የቅድመ ስልጠና ዝቅተኛ ገደብ ምስጢር አይደለም — የማሽን መማር ለአሥርተ ዓመታት የተረዳው ተራ የስታቲስቲክስ ስህተት ነው። መቀጠሉም የግድ አይደለም — በከፊል እኛ የገነባነውና ልንቀይረው የምንችለው ማበረታቻ ነው። እርግጠኛ ሳይሆን መልስ መስጠትን የሚተው ሥርዓት ምንም ቅዠት አያደርግም፤ ተግባራዊ ሞዴሎች ይህን የማያደርጉት leaderboards መቆጠብን ስለሚቀጡ ነው።

ደራሲዎቹ ምን አደረጉ?

ጽሑፉ ሦስት ክፍሎች አሉት። መጀመሪያ፣ “ትክክለኛ ጽሑፍ ብቻ አመንጭ” የሚለው “ይህ መግለጫ ትክክል ነው?” ከሚለው ሁለትዮሽ የመመደብ ችግር ቢያንስ እንደሚከብድ በማሳየት፣ አንዳንድ ቅጥት በቅድመ ስልጠና ጊዜ በስታቲስቲክስ እንደሚገደድ የሂሳብ ክርክር ያቀርባል። ሁለተኛ፣ አሥር ተፅዕኖ ያላቸውን benchmarks በመመርመር፣ የተለመዱ የaccuracy መለኪያዎች ከመቆጠብ ይልቅ ግምትን እንደሚሸልሙ ይከራከራል። ሦስተኛ፣ የቀረበ መፍትሔና የሚፈትነው የጉዳይ ጥናት አለ፦ open-rubric ግምገማዎች፣ ውጤቱ በጥያቄው ውስጥ ራሱ የሚገለጽባቸው (ለምሳሌ፣ “ትክክለኛ መልስ 1፣ የተሳሳተ መልስ −1 ያገኛል፣ ስለዚህ እርግጠኝነትዎ ከ50% በታች ከሆነ ይቆጠቡ”)፣ ሞዴሉም ታማኝነት መቼ እንደሚሸለም እንዲያውቅ። ይህን በአራት ግንባር ቀደም ሞዴሎች — የGoogle Gemini 3 Pro፣ የOpenAI GPT-5፣ የxAI Grok 4 እና የAnthropic Claude Opus 4.5 — ላይ፣ 4,326 የSimpleQA የእውነታ ጥያቄዎችን በመጠቀም ሞክረዋል። የጉዳይ ጥናቱ ለማሳየት ብቻ እንደሆነ፣ “በሞዴሎች መካከል ቁጥጥር ያለው ግምገማ እንዳልሆነ” በግልጽ ይናገራሉ (መደበኛ ቅንብሮች፣ ማስተካከያ የለም፣ ወጪ አልተመጣጠነም)።

ምን አገኙ?

- **ቅድመ ስልጠና አንዳንድ ስህተትን ያስገድዳል።** ሞዴሉ በእርግጠኝነት ሐሰት የሚያመነጭበት መጠን፣ ከእሱ የተገነባው ምርጥ “ይህ መግለጫ ትክክል ነው?” መመደቢያ ከሚኖረው የስህተት መጠን (በግምት ሁለት እጥፍ) የሚመነጭ ዝቅተኛ ገደብ አለው። ሊማር የሚችል ሥርዓት ለሌላቸው እውነታዎች፣ ይህ ገደብ ቢያንስ *singleton* መጠን — በስልጠና ውስጥ አንድ ጊዜ ብቻ የሚታዩ እውነታዎች ድርሻ — ነው። መረጃው ፍጹም ንጹሕ ቢሆንም አንዳንድ ቅጥት የማይቀር ነው።
- **ውጤት አሰጣጥ ግምትን ይሸልማል — በተግባር።** በተለመደ ትክክል/ስህተት አሰጣጥ ፈጽሞ አለመቆጠብ ምርጥ ስልት ነው፣ እና የደራሲዎቹ ጥናት አብዛኞቹ ታዋቂ benchmarks “አላውቅም”ን እንደ ተሳሳተ መልስ እንደሚቆጥሩ አገኘ። ከራሳቸው ጎን ግልጽ ምሳሌ፦ በSimpleQA ፈተና፣ ቀጥተኛ accuracy ሁሉን የሚመልስና ከጊዜው ከሦስት አራተኛ በላይ የሚሳሳተውን የOpenAI o4-mini፣ እርግጠኛ ሳይሆን ስለሚቆጠብ እጅግ ያነሰ የሚሳሳተውን GPT-5-mini በትንሹ ይበልጥ ይመርጣል። ይበልጥ ደፋር ሞዴል በleaderboard የተሻለ ይታያል።
- **Open rubrics ማበረታቻውን ይገለብጣሉ (በጉዳይ ጥናታቸው)።** ቀላል ቅጥት መቀነሻን ፈትነዋል (ሞዴሉ ሁለት ጊዜ ይመልስ፣ መልሶቹ ካልተስማሙ ይቆጠብ)። በተለመደ accuracy፣ ዘዴው ስህተትን ይቀንሳል ግን *accuracy*ንም ይቀንሳል — ስለዚህ መለኪያው መጠቀሙን አያበረታታም። በopen rubrics ግን ያው ዘዴ በተለያዩ የቅጣት ደረጃዎች ለአራቱም ሞዴሎች ይሻላል፤ ቀጥተኛ accuracy እርግጠኛ ሳይሆን በመቆጠቡ የቀጣው GPT-5-miniም ውጤቱ በግልጽ ከተነገረ በኋላ o4-miniን ይበልጣል (n = 4,326 ጥያቄዎች ለእያንዳንዱ ሞዴል)።

ይህ ምን ማለት ይሆናል?

ቅጥትን መቀነስ በዋናነት ተጨማሪ ልዩ ፈተናዎችን መፍጠር አይደለም። “አላውቅም” ማለት እንዳይቀጣ፣ ዋና benchmarks ጥርጣሬን የሚመዝኑበትን መንገድ መቀየር ነው። Leaderboard እስካልተቀየረ፣ ቅጥትን መቀነስ ሞዴሎችን accuracy ነጥብ ማስከፈሉንና ስለዚህ አለመበረታታቱን ይቀጥላል — ደራሲዎቹ ችግሩን “ማህበራዊ-ቴክኒካዊ” የሚሉት ለዚህ ነው፤ አንድ ክፍል የተሻለ መለኪያ፣ ሌላው ክፍል ተፅዕኖ ያላቸው leaderboards እንዲቀበሉት ማድረግ ነው።

ይህ የማያረጋግጠው

- Open rubrics በተግባራዊ ዓለም ቅዠትን እንደሚፈቱ **አያሳይም**። የድጋፍ ሙከራው ትንሽና ሆን ተብሎ **ቁጥጥር የሌለው** የጉዳይ ጥናት ነው — አራት ሞዴሎች በመደበኛ ቅንብር፣ አንድ የተመረጠ መቀነሻ፣ አንድ የእውነታ ጥያቄ ፈተና — የማበረታቻ መገልበጥን ለማሳየት እንጂ ሞዴሎችን ለመደርደር ወይም አጠቃላይ ውጤታማነትን ለማረጋገጥ አይደለም።
- ውጤት አሰጣጥ ብቸኛው መንስኤ እንደሆነ **አይናገርም**። በስልጠና መረጃ ውስጥ ያሉ ስህተቶች፣ በእውነት ከባድ ችግሮች እና ያልተለመዱ prompts ሌሎች ምንጮች ሆነው ይቀራሉ።
- ቅዠት የማይቀር ነው የሚለውን ታዋቂ ንግግር **አይደግፍም**። ደራሲዎቹ ተቃራኒውን ይከራከራሉ፤ ሊረጋገጡ የሚችሉ ጥያቄዎችን ብቻ የሚመልስና ሌላውን "አላውቅም" የሚል ሥርዓት ፈጽሞ ቅዠት አያደርግም።
- የቅድመ ስልጠና ዝቅተኛ ገደብን **አያጠፋም** — ያብራራዋል፣ ይገድበዋል፤ ገደቡም በእርግጠኝነት የተነገሩ የእውነታ ስህተቶችን እንጂ ሁሉንም የሞዴል ባህሪ አይመለከትም።
- Open rubrics ብቻቸውን *በቂ* እንደሆኑ **አያሳይም**። ግምገማው የሚሸልመውን ይቀይራሉ፤ retrieval፣ የመሣሪያ አጠቃቀም ወይም የተሻለ የጥርጣሬ መጠን ያላቸው ሞዴሎችን አይተኩም።

ማስረጃው ምን ያህል ጠንካራ ነው?

- ዋናው **ሂሳብ** ነው — መደበኛ ዝቅተኛ ገደቦች፣ መለኪያዎች አይደሉም። እንደ ቲዎሪ ክርክር በራሱ ሁኔታ ጠንካራ ነው።
- ችግሩን **ሆን ብሎ በሚያቀልሉ ሞዴሎች** ላይ ይመሠረታል፤ ደራሲዎቹ እያንዳንዱን መልስ ትክክል፣ ስህተት ወይም "አላውቅም" ብሎ የሚከፍለውን "false trichotomy" እና ለግልጹ ገደብ የተጠቀሙበትን ሃሳባዊ "የዘፈቀደ እውነታዎች" ሁኔታ ይጠቅሳሉ።
- የbenchmark ጥናቱ **ትንሽ የተመረጠ ናሙና** ነው — አሥር ተፅዕኖ ያላቸው ግምገማዎች፣ ሙሉ ምርመራ አይደለም።
- የጉዳይ ጥናቱ **እውነተኛ ግን ውስን** ነው፤ አራት ግንባር ቀደም ሞዴሎች፣ አንድ መቀነሻ፣ SimpleQA ብቻ፣ መደበኛ ቅንብሮች፣ እና "ቁጥጥር ያለው ግምገማ አይደለም" ተብሎ በግልጽ ተገልጿል። ለማበረታቻ ክርክር የፅንሰ ሐሳብ ማረጋገጫ ነው፣ benchmark ውጤት አይደለም።
- የሐሳቡን የመጣበት ቦታ መጥቀስ ይገባል፤ ከአራቱ ደራሲዎች ሦስቱ በOpenAI ይሠራሉ ወይም ሠርተው ነበር፣ እና ጽሑፉ መስኩ ሞዴሎችን የሚገመግምበትን መንገድ እንዲቀይር ይከራከራል። ይህ ጥቅም ካለው ወገን የመጣ ጥሩ ክርክር ነው፣ ገለልተኛ ውጫዊ ግምገማ አይደለም — ሊመዘን እንጂ ሊጣል አይገባም። (ጽሑፉ ትችቱን ለራሱ o4-mini እና GPT-5-mini ሞዴሎችም እንደ ሌሎች ስለሚያደርግ ይመሰገናል።)

ለምን አስፈላጊ ነው?

በጣም የተጋነነ ችግርን እንደገና ያቀርባል። "ቅዠት" እንደ አስፈሪ ጉድለት ወይም የማይንቀሳቀስ ግድግዳ ይቀርባል፤ ይህ ጽሑፍ ግን ተራና በከፊል በራሳችን የተፈጠረ ያደርገዋል — ልንረዳው የምንችለው የስታቲስቲክስ ዝቅተኛ ገደብ፣ እኛ በመረጥነው ማበረታቻ ላይ ተቀምጦ። ሰፊው ትምህርት የተረጋጋና የበለጠ ጠቃሚ ነው፡- በአስተማማኝነት ላይ ያለው ቀጣይ እድገት በምንገነባው ያህል *በምንለካው* ላይ ሊመሠረት ይችላል።

ግልጽ ማጠቃለያ

የቋንቋ ሞዴሎች በእርግጠኝነት የሚሰጧቸው ሐሰተኛ መልሶች ከሁለት ቦታዎች ይመጣሉ። የመጀመሪያው ስታቲስቲካዊ ነው፤ እውነታ ሊማር የሚችል ሥርዓት ከሌለው፣ ቋንቋን ለመኮረጅ የሰለጠነ ሞዴል አንዳንድ ጊዜ ይሳሳታል፣ ይህ ዝቅተኛ ገደብም ሊገመት ይችላል (ለምሳሌ፣ ከአንድ ጊዜ ብቻ ከሚታዩ እውነታዎች)። ሁለተኛው ማበረታቻ ነው፤ ሞዴሎች የሚደረደሩባቸው ከሞላ ጎደል ሁሉም benchmarks “አላውቅም”ን ከተሳሳተ መልስ ጋር እኩል ይመዝናሉ፣ ስለዚህ ግምት ሁልጊዜ ያሸንፋል — እስከ ሦስት አራተኛ ጊዜ የሚሳሳት ሞዴል በታማኝነት ከሚቆጠብ ሞዴል ሊበልጥ ድረስ። የደራሲዎቹ ሐሳብ ሌላ ቅጥት ፈተና ሳይሆን “open rubrics” ነው፦ ውጤት አሰጣጡን በጥያቄው ውስጥ ይግለጹ። በአራት ግንባር ቀደም ሞዴሎች የጉዳይ ጥናት፣ ይህ ማበረታቻውን ይገለብጣል፣ ቅጥትን የሚቀንስ ዘዴም ከመቀጣት ይልቅ ይሸለማል። ይህ በ*Nature* የተገመገመ፣ ቲዎሪንና ጥናትን ከትንሽ እና በግልጽ ቁጥጥር ከሌለው ሙከራ ጋር ያጣመረ ጽሑፍ ነው፤ መፍትሔው ተስፋ ይሰጣል ግን በስፋት እንደሚሠራ ገና አልታየም፣ ቅጥትም ምስጢራዊ ወይም ፍጹም የማይቀር እንዳልሆነ ይከራከራል።

ከማጋነን ነፃ ምርመራ

ጽሑፉ የሚያሳየው: በቅድመ ስልጠና አንዳንድ ቅጥትን የሚያስገድድ የሂሳብ ዝቅተኛ ገደብ (ሥርዓት ለሌላቸው እውነታዎች ቢያንስ “singleton መጠን”)፤ አብዛኞቹ ዋና benchmarks “አላውቅም” ለሚለው ነጥብ እንደማይሰጡ ያሳየ ጥናት፤ እና ውጤትን በprompt ውስጥ መግለጽ (“open rubrics”) ቅጥትን የሚቀንስ ዘዴ እንዲያሸንፍ ያደረገበት፣ ቀጥተኛ accuracy ግን የቀጣው የአራት ሞዴል ጥናት።

ሊሆን የሚችለው ግን ያልተረጋገጠው: Open rubrics በዋና benchmarks ውስጥ ቢገቡ፣ በተግባራዊ ሞዴሎች ቅጥትን በትርጉም ያለው መጠን እንደሚቀንሱ። የድጋፍ ሙከራው ትንሽና ቁጥጥር እንደሌለው በግልጽ ተገልጿል።

የማያሳየው: ቅጥት የማይቀር መሆኑን (ተቃራኒውን ይከራከራል)፤ ውጤት አሰጣጥ ብቸኛው መንስኤ መሆኑን፤ ቅጥት ሙሉ በሙሉ ሊጠፋ እንደሚችል፤ ወይም የጉዳይ ጥናቱ አራቱን ሞዴሎች እንደሚደረድር።

ዋና ገደቦች: ሆን ተብሎ የቀለለ ትክክል/ስህተት/“አላውቅም” ሞዴል (ደራሲዎቹ “false trichotomy” ይሉታል)፤ ትንሽ የተመረጠ benchmark ጥናት (አሥር ግምገማዎች)፤ ቁጥጥር የሌለው የጉዳይ ጥናት (አራት ሞዴሎች፣ አንድ ዘዴ፣ አንድ ፈተና፣ መደበኛ ቅንብሮች)፤ እና OpenAI የመራው መስኩ ሞዴሎችን እንዴት ሊገመግም እንደሚገባ የሚያቀርብ ክርክር።

አጠቃላይ አንባቢ ምን ያህል መተማመን ይኖርበታል? ቅጥት ምስጢራዊ ወይም ፍጹም የማይቀር እንዳልሆነ፣ እና ዋና benchmarks አሁን ግምትን እንደሚሸልሙ ከፍተኛ መተማመን። የቀረበው መፍትሔ እንደሚረዳ መካከለኛ መተማመን — አሁን እውነተኛ የፅንሰ ሐሳብ ማረጋገጫ አለው፣ በስፋትና በትልቅ ደረጃ እንደሚሠራ ግን ገና አልታየም።

ምንጮች

Nature 653, 1047–1050 (2026)
https://doi.org/10.1038/s41586-026-10549-w

Why Language Models Hallucinate (arXiv 2509.04664)
https://arxiv.org/abs/2509.04664

hallucinations-paper-experiments (OpenAI)
https://github.com/openai/hallucinations-paper-experiments

SimpleQA
https://github.com/openai/simple-evals