The Clean Paper

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

Kwa nini modeli za lugha hutoa hallucination — na kwa nini namna tunavyozipa alama huendeleza hali hiyo

Uongo unaotolewa kwa kujiamini ambao tunauita "hallucination" si hitilafu ya kifumbo: baadhi yake ni matokeo ya takwimu ya mafunzo, na huendelea kwa sababu vipimo vikuu hutuza kisio la kujiamini kuliko "sijui" ya uaminifu. Uchunguzi kifani wa modeli nne za kiwango cha juu unaonyesha kwamba kutaja kanuni za alama ndani ya maelekezo ("kanuni wazi") hugeuza motisha hiyo.

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

Kwa nini modeli hukisia, na kwa nini tulizifundisha kufanya hivyo

Uliza modeli kubwa ya lugha tarehe ya kuzaliwa ya mtu usiyemjua, na inaweza kujibu “Machi 7” kwa uhakika thabiti wa mtu anayesoma jibu kwenye kadi — lakini ikosee mara tatu mfululizo kwa kutoa tarehe tatu tofauti. Waandishi wanatoa mfano wa aina hiyo hasa: modeli maarufu ziliulizwa swali rahisi la ukweli — tarehe ya kuzaliwa ya mtu au maana kamili ya kifupisho kisichojulikana — na kila moja ikabuni jibu tofauti kwa kujiamini, bila hata moja kuwa sahihi. Sekta huita hili *hallucination*, neno linalofanya lionekane kama hitilafu ya utambuzi. Hatua ya kwanza ya makala ni kuondoa fumbo hilo.

Anza na jinsi modeli inavyoundwa. Katika hatua yake ya kwanza na kubwa zaidi ya mafunzo, hujifunza kwa vitendo lugha yenye mtiririko inaonekanaje kwa kusoma kiasi kikubwa sana cha maandishi. Sasa fikiria ukweli usio na muundo wa kuufuata — tarehe ya kuzaliwa ya mtu fulani. Ikiwa tarehe hiyo ilionekana mara moja tu katika maandishi ya mafunzo, au haikuonekana kabisa, hakuna kitu ambacho mfumo unaojifunza mifumo unaweza kushika: kwa mtazamo wa modeli, jibu ni la kubahatisha. Waandishi wanaweka hoja hii kwa usahihi kwa kukopa wazo la zamani (la Alan Turing, kutoka tatizo tofauti): ikiwa tarehe moja kati ya tano za kuzaliwa inaonekana mara moja tu katika data, modeli inapaswa kutarajiwa kukosea angalau moja kati ya tano — si kwa sababu imeharibika, bali kwa sababu hakukuwa na kitu cha kujifunza. (Kwa mantiki hiyo hiyo, modeli karibu kamwe hazikosei mji mkuu wa nchi: taarifa hizo huonekana mara kwa mara.) Wanatoa hoja kwa tahadhari kwamba kutofautisha taarifa ya kweli na ya uongo lakini inayosikika kuwa ya kweli ni tatizo gumu, na kutoa taarifa za kweli *pekee* ni angalau kugumu kwa kiwango hicho. Kiwango fulani cha chini cha makosa kimejengeka ndani ya mfumo.

Wazo la msingi: jinsi ya kuhesabu kile ambacho bado hujakiona

Hoja hii inategemea wazo lenye werevu kweli — la zamani kuliko modeli za lugha na linalostahili kueleweka vizuri.

Anza na mfuko wa mipira yenye rangi. Hujui una rangi ngapi. Unatoa mipira 100, mmoja baada ya mwingine, na kuihesabu: myekundu 40, bluu 25, kijani 15, njano 5, zambarau 3, chungwa 2 — kisha **rangi kumi tofauti ambazo kila moja inaonekana mara moja tu.**

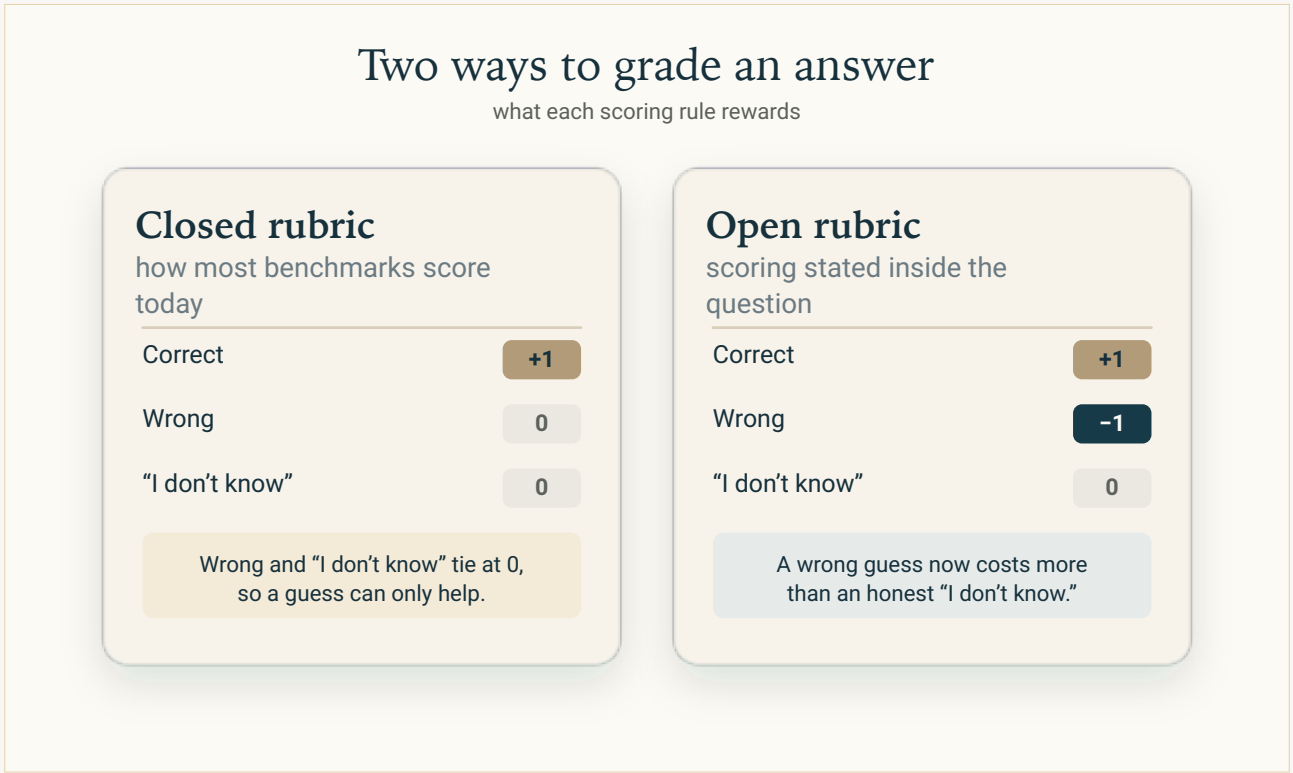
Sasa fikiria swali ambalo Turing alikabili kwa kweli, katika tatizo tofauti kabisa: *kuna uwezekano gani kwamba mpira unaofuata utakuwa wa rangi ambayo hujawahi kuona?* Huwezi kuhesabu kitu ambacho hujawahi kutoa — lakini **unaweza** kuhesabu rangi ulizoona mara moja tu, zinazoitwa “singletons.” Mbinu inayoitwa **ukadiriaji wa Good–Turing** hutumia sehemu ya mipira iliyotolewa ambayo ni singleton kukadiria uwezekano ambao bado umejificha katika rangi ambazo hujaona. Mipira kumi kati ya mia moja ulizotoa ilikuwa ya rangi iliyoonekana mara moja tu, kwa hiyo uwezekano wa mpira unaofuata kuwa wa rangi mpya kabisa ni karibu **10 / 100 = 10%.**

Rangi hizo zilizoonekana mara moja si *makosa*. Ni kipimo cha ujinga wako mwenyewe: rangi nyingi zinapoonekana mara moja, sampuli inakuambia kwamba ulimwengu una nyingine ambazo bado hujazitoa.

Sasa badilisha rangi ziwe tarehe za kuzaliwa, na mfuko uwe maandishi ya mafunzo ya

> modeli. Tuseme kwamba, kati ya tarehe ilizoona, **moja kati ya tano inaonekana mara moja tu.** Mbinu ni ileile: karibu sehemu moja ya tano ya uwezekano iko katika tarehe ambazo kwa vitendo modeli haijawahi kuona — na tarehe ya kuzaliwa haina muundo wa kutegemea (huwezi *kukokotoa* tarehe ya kuzaliwa ya mtu). Kwa hiyo tarehe iliyoonekana mara moja au kutowahi kuonekana ni kama sarafu ambayo modeli haiwezi kuipa uzito sahihi, na itakosea karibu **moja kati ya tano**. Hakuna ujanja unaoweza kurekebisha hili: hakukuwa na kitu cha kujifunza.
> Hiyo ndiyo hoja yote kwa ufupi: kiwango cha singleton hupima kiasi cha ulimwengu ambacho hakiwezi kujifunzwa *kutokana na data hii*, na hicho huwa **kiwango cha chini** cha makosa. Pia ndicho kinachofanya modeli karibu kamwe isikose mji mkuu — *Paris* huonekana mara kwa mara, kiwango chake cha singleton kiko karibu na sifuri, hivyo kuna mengi ya kujifunza.

Hilo linaeleza hallucination zinatoka wapi. Halielezi kwa nini *zinaendelea kuwepo* — kwa nini modeli, baada ya mafunzo yote ya baadaye yaliyokusudiwa kuzifanya ziwe na msaada na uaminifu, bado hujifanya zinajua badala ya kukiri shaka. Mfano wa makala hapa unafaa kwa namna inayokera kidogo. Fikiria mwanafunzi katika mtihani ambaye hajui jibu. Ikiwa kuacha nafasi tupu kunapata sifuri na kukisia kunaweza kupata alama moja, njia ya kuongeza alama ni kukisia — kwa kujiamini na kwa usahihi wa kuigiza, kamwe si “sina uhakika.” Wanafunzi hujifunza hilo. Inageuka kuwa modeli pia hujifunza, kwa sababu tunazipa alama kwa njia hiyo hiyo. Waandishi walichunguza vipimo sanifu ambavyo taaluma hushindania na jedwali za uongozi ambazo modeli hurekebishwa ili zipande, na wakagundua kwamba karibu vyote huipa “sijui” alama sawa kabisa na jibu lisilo sahihi: sifuri. Chini ya kanuni hiyo, modeli inayokisia kila wakati itaishinda modeli inayofanana nayo lakini inayotaja kutokuwa na uhakika kwa uaminifu. Kwa maana iliyo karibu kabisa na halisi, tunazipa alama zinazozisukuma kufanya hivyo.



Vipimo sanifu vinaweza kufanya kukisia kuwe na mantiki. Kwa mfumo wa alama unaotumiwa na vipimo vingi (kushoto), jibu lisilo sahihi na “sijui” ya uaminifu zote hupata sifuri — hivyo kukisia kunaweza kusaidia tu. Kanuni zikiwekwa ndani ya swali na kukawa na adhabu kwa kukosea (kulia), kukataa kujibu ukiwa na shaka kunaweza kuwa uamuzi bora. Hilo hubadilisha kile ambacho jaribio linatuza; peke yake halitatui hallucination.

Hii ndiyo sehemu inayostahili kubaki akilini kwa sababu inaenda kinyume na kichwa cha habari cha kawaida. Hallucination mara nyingi huuzwa kama kikomo kisichoepukika na karibu cha kifumbo cha teknolojia. Makala inapinga mawazo yote mawili. Kiwango cha chini kinachotokana na mafunzo ya awali si fumbo — ni kosa la kawaida la takwimu ambalo ujifunzaji wa mashine umelielewa kwa miongo kadhaa. Na kuendelea kwake si jambo lisiloepukika — kwa sehemu ni motisha tuliyoijenga na tunayoweza kuibadilisha. Mfumo ambao ungeamua tu kutokujibu ukiwa na shaka usingetoa hallucination hata moja; sababu modeli zinazotumika hazifanyi hivyo ni kwamba jedwali zetu za alama huadhibu kukataa.

Waandishi walifanya nini

Makala ina sehemu tatu. Kwanza, kuna hoja ya kihisabati kwamba baadhi ya hallucination hulazimishwa na takwimu wakati wa mafunzo ya awali, kwa kuonyesha kwamba “toa maandishi halali pekee” ni angalau kugumu sawa na tatizo la uainishaji wa jozi la “taarifa hii ni halali?” Pili, kuna hoja — inayoungwa mkono na uchunguzi wa vipimo kumi vyenye ushawishi — kwamba vipimo vya kawaida vinavyotegemea usahihi hutuza kukisia kuliko kukataa kujibu. Tatu, kuna suluhisho lililopendekezwa na uchunguzi kifani unaolijaribu: tathmini zenye **kanuni wazi za alama**, ambapo namna ya kutoa alama inaelezwa ndani ya swali lenyewe (kwa mfano, “jibu sahihi linapata 1, lisilo sahihi −1, kwa hiyo kataa kujibu ikiwa una uhakika wa chini ya 50%”), ili modeli iweze kujua wakati uaminifu unatuza. Wanajaribu hilo kwenye modeli nne za kiwango cha juu — Gemini 3 Pro ya Google, GPT-5 ya OpenAI, Grok 4 ya xAI na Claude Opus 4.5 ya Anthropic — kwa kutumia maswali 4,326 ya ukweli ya SimpleQA. Wanasema wazi kwamba uchunguzi huo ni wa kuonyesha mfano, “si tathmini inayodhibitiwa kati ya modeli” (mipangilio ya kawaida, hakuna urekebishaji, hakuna kusawazisha gharama).

Waligundua nini

- **Mafunzo ya awali hulazimisha baadhi ya makosa.** Kiwango ambacho modeli hutoa uongo kwa kujiamini kina mpaka wa chini unaotokana na (karibu mara mbili ya) kiwango cha kosa cha mfumo bora wa uainishaji wa “taarifa hii ni halali?” uliojengwa kutokana nayo. Kwa mambo yasiyo na muundo unaoweza kujifunzwa, mpaka huo ni angalau *kiwango cha singleton* — sehemu ya mambo yanayoonekana mara moja tu katika mafunzo. Baadhi ya hallucination haiwezi kuepukwa hata data ikiwa safi kabisa.
- **Utoaji alama hutuza kukisia — kwa hali halisi.** Chini ya alama za kawaida za sahihi/kosa, kutokataa kujibu kamwe ndiyo mikakati bora, na uchunguzi wa waandishi unaonyesha kuwa idadi

kubwa ya vipimo maarufu huchukulia “sijui” kuwa kosa tu. Mfano wazi kutoka kwao wenyewe: katika jaribio la SimpleQA, usahihi ghafi unaipendelea kidogo o4-mini ya OpenAI — ambayo hujibu karibu kila kitu na hukosea zaidi ya robo tatu ya wakati — kuliko GPT-5-mini, ambayo hufanya makosa machache sana kwa sababu hukataa kujibu ikiwa haina uhakika. Modeli isiyo na tahadhari inaonekana bora kwenye jedwali la alama.

- **Kanuni wazi hugeuza motisha (katika uchunguzi wao kifani).** Wanajaribu njia rahisi ya kupunguza hallucination (modeli ijibu mara mbili na ikatae ikiwa majibu mawili hayatani). Chini ya usahihi wa kawaida, njia hiyo hupunguza makosa *lakini pia hupunguza usahihi* — hivyo kipimo kinakatisha tamaa ya kuitumia. Chini ya kanuni wazi, njia hiyohiyo hushinda kwa modeli zote nne katika viwango mbalimbali vya adhabu; na GPT-5-mini — ambayo usahihi ghafi uliiadhibu kwa kukataa kujibu ikiwa haina uhakika — huipita o4-mini baada ya mfumo wa alama kuelezwa wazi (n = 4,326 maswali kwa kila modeli).

Hili labda linamaanisha nini

Kupunguza hallucination kwa kiasi kikubwa si suala la kubuni majaribio zaidi maalumu kwa hallucination. Ni suala la kubadilisha jinsi vipimo vikuu vinavyotoa alama kwa kutokuwa na uhakika, ili kukiri “sijui” kusiadhibiwe tena. Hadi jedwali la alama libadilike, kupunguza hallucination kutaendelea kugharimu modeli alama za usahihi na hivyo kuendelea kukatishwa tamaa — ndiyo maana waandishi wanaliita tatizo la “kijamii na kiteknolojia”: sehemu moja ni kipimo bora, na sehemu nyingine ni kupata jedwali zenye ushawishi zikikubali kipimo hicho.

Hili *halithibitishi* nini

- **Halionyeshi** kwamba kanuni wazi za alama hurekebisha hallucination katika matumizi ya kawaida. Jaribio linalounga mkono hoja ni uchunguzi kifani mdogo na kwa makusudi **usiomedhibitiwa** — modeli nne kwa mipangilio ya kawaida, njia moja iliyochaguliwa ya kupunguza tatizo, na jaribio moja la maswali ya ukweli — unaokusudiwa kuonyesha *mgeuko wa motisha*, si kupanga modeli kwa ubora au kuthibitisha ufanisi wa jumla.
- **Halidai** utoaji alama ndiyo sababu *pekee*. Makosa katika data ya mafunzo, matatizo magumu kwa kweli na maelekezo yasiyozoeleka hubaki vyanzo tofauti.
- **Haliungi mkono** kauli maarufu kwamba hallucination haziepukiki. Waandishi wanatoa hoja ya kinyume: mfumo unaojibu tu maswali yanayoweza kukaguliwa na kusema “sijui” kwa mengine usingetoa hallucination kamwe.
- **Haliondoi** kiwango cha chini cha makosa ya mafunzo ya awali — linakieleza na kuweka mipaka, na mpaka huo unahusu makosa ya ukweli yanayotolewa kwa kujiamini, si tabia zote za modeli.
- **Halionyeshi** kwamba kanuni wazi *zinatosha* peke yake. Zinabadilisha kile ambacho tathmini inatuza; si mbadala wa urejeshaji wa taarifa, matumizi ya zana au modeli zenye kipimo bora cha uhakika.

Ushahidi una nguvu kiasi gani?

- Msingi ni **hisabati** — mipaka rasmi ya chini, si vipimo vya majaribio. Kama hoja ya kinadharia, ina nguvu ndani ya masharti yake.
- Inategemea **modeli zilizorahisishwa kwa makusudi** za tatizo; waandishi wenyewe wanataja "mgawanyo wa uongo wa aina tatu" unaochukulia kila jibu kuwa sahihi, lisilo sahihi au "sijui," pamoja na hali bora iliyobuniwa ya "mambo ya kubahatisha" inayotumiwa kupata mpaka ulio wazi zaidi.
- Uchunguzi wa vipimo ni **sampuli ndogo iliyochaguliwa** — tathmini kumi zenye ushawishi, si ukaguzi wa kila kitu.
- Uchunguzi kifani ni **halisi lakini wenye mipaka**: modeli nne za kiwango cha juu, njia moja, SimpleQA pekee, mipangilio ya kawaida, na umeelezwa wazi kuwa "si tathmini inayodhibitiwa." Ni uthibitisho wa dhana kuhusu motisha, si matokeo ya kulinganisha modeli.
- Inafaa kutaja mtazamo wa chanzo: waandishi watatu kati ya wanne wanafanya au waliwahi kufanya kazi katika **OpenAI**, na makala inatoa hoja kwamba taaluma ibadilishe jinsi inavyotathmini modeli. Huo ni msimamo wenye hoja kutoka kwa upande wenye maslahi, si mapitio huru ya nje — unapaswa kupimwa, si kukataliwa. (Kwa upande mzuri, makala inaelekeza ukosoaji kwa modeli zake mwenyewe, o4-mini na GPT-5-mini, sawa na inavyofanya kwa nyingine.)

Kwa nini ni muhimu

Inabadilisha namna ya kuangalia tatizo lililopambwa sana. "Hallucination" huuzwa ama kama kasoro ya kutisha au ukuta usiosogezeka; makala hii inalifanya kuwa la kawaida na kwa sehemu tulilojisababishia — kiwango cha chini cha takwimu tunachoweza kuelewa, kikiwa juu ya motisha tuliyochagua. Funzo pana ni tulivu na lenye manufaa zaidi: maendeleo ya kuaminika yanaweza kutegemea *kile tunachopima* kwa kiwango sawa na kile tunachojenga.

Muhtasari safi

Majibu ya uongo yanayotolewa kwa kujiamini na modeli za lugha hutoka sehemu mbili. Ya kwanza ni ya takwimu: jambo lisipokuwa na muundo wa kujifunza, modeli iliyofunzwa kuiga lugha itakosea wakati mwingine, na kiwango hicho cha chini kinaweza kukadiriwa (kwa mfano, kutokana na idadi ya mambo yanayoonekana mara moja tu katika mafunzo). Ya pili ni motisha: karibu kila kipimo ambacho modeli hupangwa nacho hupa "sijui" alama sawa na jibu lisilo sahihi, hivyo kukisia hushinda kila wakati — kiasi kwamba modeli inayokosea robo tatu ya wakati inaweza kuipita modeli yenye uaminifu zaidi inayokataa kujibu. Pendekezo la waandishi si jaribio jingine la hallucination bali "kanuni wazi": eleza mfumo wa alama ndani ya swali. Katika uchunguzi kifani wa modeli nne za kiwango cha juu, hilo hugeuza motisha ili njia ya kupunguza hallucination ituzwe badala ya kuadhibiwa. Hii ni makala ya nadharia pamoja na uchunguzi, yenye jaribio dogo lisilodhibitiwa lililotajwa wazi, na ilipitiwa na wataalamu katika *Nature*; suluhisho lina matumaini lakini bado halijaonyeshwa kufanya kazi kwa

kiwango kikubwa, na hoja ni kwamba hallucination si fumbo wala jambo lisiloepukika kabisa.

Ukaguzi usio na upotoshaji

Makala inaonyesha nini: Mpaka wa chini wa kihisabati ambao chini yake baadhi ya hallucination hulazimishwa wakati wa mafunzo ya awali (angalau “kiwango cha singleton” kwa mambo yasiyo na muundo); uchunguzi unaoonyesha kwamba vipimo vingi maarufu havipi “sijui” alama; na uchunguzi kifani wa modeli nne ambapo kutaja mfumo wa alama ndani ya maelekezo (“kanuni wazi”) hufanya njia ya kupunguza hallucination ishinde, ilhali usahihi wa kawaida uliiadhibu.

Kinachowezekana lakini hakijathibitishwa: Kwamba kanuni wazi zikijumuishwa katika vipimo vikuu zitapunguza hallucination kwa kiwango cha maana katika modeli zinazotumika. Jaribio linalounga mkono hoja ni dogo na lisilodhibitiwa kwa uwazi.

Kisichoonyeshwa: Kwamba hallucination haziepukiki (makala inatoa hoja ya kinyume); kwamba utoaji alama ndiyo sababu pekee; kwamba hallucination inaweza kuondolewa moja kwa moja; au kwamba uchunguzi kifani unapanga modeli nne dhidi ya nyingine.

Mipaka mikuu: Modeli iliyorahisishwa kwa makusudi ya sahihi/kosa/“sijui” (waandishi wanaiita “mgawanyo wa uongo wa aina tatu”); uchunguzi mdogo uliochaguliwa wa vipimo (tathmini kumi); uchunguzi kifani usiodhibitiwa (modeli nne, njia moja, jaribio moja, mipangilio ya kawaida); na hoja inayoongozwa na OpenAI kuhusu namna taaluma inavyopaswa kutathmini modeli.

Msomaji wa kawaida awe na uhakika kiasi gani? Uhakika mkubwa kwamba hallucination si fumbo wala jambo lisiloepukika kabisa, na kwamba vipimo vikuu kwa sasa hutuza kukisia. Uhakika wa wastani kwamba suluhisho lililopendekezwa linasaidia — sasa lina uthibitisho halisi wa dhana, lakini bado hakuna onyesho kuwa linafanya kazi kwa upana na kwa kiwango kikubwa.

Vyanzo

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