LLM Sozlamalari

- Harorat, Top P, va Maksimal Uzunlikni tushunish

Kirish

Biz modelning turli jihatlarini boshqarish uchun ma'lum LLM[1] sozlamalaridan foydalanishimiz mumkin, masalan, uning qanchalik 'tasodifiy' ekanligi. Ushbu sozlamalar yanada ijodkor, xilma-xil va qiziqarli natijalar olish uchun moslashtirilishi mumkin. Temperature, Top P va Max Length sozlamalari eng muhim, lekin biz OpenAI Playground sizga o'zgartirish imkonini beradigan har bir sozlamani tasvirlaymiz.

Harorat

Temperatura til modelining natijalaridagi noaniqlikni boshqaradi. Harorat sozlamalari yuqori bo'lganda, natijalar yanada ijodiy va kamroq kutilganidek, chunki u kamroq ehtimolli tokenlarning ehtimolini oshiradi, shu bilan birga ehtimoli yuqori bo'lganlari uchun kamaytirish. Aksincha, past haroratlar natija beradi ko'proq konservativ va oldindan aytib bo'ladigan natijalar. Quyidagi misol ko'rsatadi ushbu chiqishdagi farqlar:

Qumloqda qilish mumkin bo'lgan 10 ta g'alati, noyob va qiziqarli narsalar nima? Ro'yxat tuzing tavsiflarsiz.

`{ 1. Qumdan qasr qurish 2. Qobiq terish 3. Plyajda voleybol o'ynash 4. Fly a kite 5. Piknik qiling 6. Paddleboardingni sinab ko'ring 7. Frisbi o'ynang 8. Snorkeling bilan shug'ullanish 9. Sohil bo'ylab uzoq sayr qiling 10. Quyosh botishini tomosha qiling }`

1. Saqich pufakchalarini puflab, sayoz suvlarda yuring

2. Murakkab qum qal'asi haykallarini yarating
3. Improvizatsiya qilingan plyaj voleybol o‘yiniga qo‘shiling
4. Yaqin atrofdagi jarliklar bo'ylab manzarali piyoda sayr qiling
5. Odamlarni kuzating va har bir kishining plyaj bilan bog'liq hikoyasini taxmin qiling.
6. Dengiz qirg‘og‘idagi san’atingizni yaratish uchun dengiz chig‘anoqlarini to‘plang
7. Sörf qilishni o'rganing yoki boshqa suv sport turlarini sinab ko'ring
8. O'zboshimchalik bilan qum jangini boshlash
9. Mahalliy aholi kabi qirg‘oqda baliq ovlashga harakat qiling
10. Qumdan qal'a qurish tanlovini tashkil qilib, do'stona raqobatda ishtirok eting

Yuqori harorat sozlamasi bilan yaratilgan chiqish plyajda bajariladigan faoliyatlarning yanada ijodiy va xilma-xil ro'yxatini taklif etadi. Bu ijodiy yozuvlar uchun juda foydali bo'lishi mumkin.

Agar haroratni juda yuqori sozlasangiz, mantiqsiz chiqishlarni olishingiz mumkin, masalan Sponge-ball baseball uyda yugurish musobaqasini Becksmith Stein Man yaqinida boshlang Beach`.

Eng yaxshi P

Top P[2] til modellari sozlamasi bo'lib, ularning chiqishidagi tasodifiylikni boshqarishga yordam beradi. U ehtimollik chegarasini o'rnatish orqali ishlaydi va keyin ushbu chegaradan oshadigan ehtimollikka ega tokenlarni tanlaydi.

Masalan, keling, model keyingisini bashorat qiladigan misolni ko'rib chiqaylik `The cat climbed up the ___` jumlasidagi so'z. Bu jumlaga mos kelishi mumkin bo'lgan eng yaxshi beshta so'z ko'rib chiqilishi mumkin `tree` (ehtimollik 0.5), `roof` (probability 0.25), `wall` (probability 0.15), `window` (probability .07) and `carpet`, with probability of .03.

Agar biz Top P ni `.90` ga o'rnatsak, AI faqat taxminan 90% ni tashkil qiluvchi tokenlarni ko'rib chiqadi. Bizning holatimizda:

- `tree` qo'shilyapti -> hozirgacha jami `50%`.

- Keyin `roof` qo'shiladi -> jami `75%` bo'ladi.
- Keyingi navbat `wall`ga, va endi yig‘indimiz `90%`ga yetadi.

Shunday qilib, chiqishni yaratish uchun, AI ushbu uchta variantdan (`tree`, `roof`, va `wall`) birini tasodifiy tanlaydi, chunki ular barcha ehtimolliklarning taxminan ~90 foizini tashkil qiladi. Ushbu usul an'anaviy usullardan ko'ra ko'proq xilma-xil chiqishlarni ishlab chiqarishi mumkin, chunki u individual tokenlarga emas, balki yig‘ma ehtimolliklarga asoslangan holda tanlovlarni qisqartiradi.

Maksimal uzunlik

Maksimal uzunlik AI tomonidan yaratilishiga ruxsat etilgan tokenlar sonining umumiy miqdoridir. Ushbu sozlama foydalanuvchilarga uzunlikni boshqarishga imkon berishi uchun foydalidir. modelning javobi, haddan tashqari uzun yoki mavzuga aloqasiz javoblarni oldini olish. Uzunlik Playground qutisidagi `USER` kiritish va`{" "}` ortasida bo'linadi `ASSISTANT` generated response. Notice how with a limit of 256 tokens, our PirateGPT from earlier is forced to cut its story short mid-sentence.

Bu, shuningdek, modeldan foydalanish uchun to'lov qilayotgan bo'lsangiz, xarajatlarni nazorat qilishga yordam beradi API o'rniga Playground'dan foydalanish.

Boshqa LLM Sozlamalari

Til modeli chiqishini ta'sir qilishi mumkin bo'lgan boshqa ko'plab sozlamalar mavjud, masalan, to'xtash ketma-ketliklari va chastota va mavjudlik jarimalari.

To'xtatish ketma-ketliklari

To'xtatish ketma-ketliklari modelga chiqishni yaratishni qachon to'xtatishni aytadi, bu esa imkon beradi sizga kontent uzunligi va tuzilishini boshqarish imkonini beradi. Agar siz AI ni yordamida sorov berayotgan bo'lsangiz email yozing, "Best regards," yoki "Sincerely," ni to'xtatish ketma-ketligi sifatida o'rnating modelni yopilish salomidan oldin to'xtashini ta'minlaydi,

bu esa elektron pochtani saqlab qoladi qisqa va aniq. To'xtatish ketma-ketliklari siz kutgan chiqish uchun foydalidir elektron pochta, raqamlangan ro‘yxat kabi tuzilgan formatda chiqish uchun yoki dialogue.

Chastota jarimasi

Chastota jarimasi yaratilgan matnda takrorlanishni oldini olish uchun sozlama hisoblanadi. matnni ularning paydo bo'lish chastotasiga nisbatan proporsional ravishda jazolab. The matn ichida token ko'proq ishlatilsa, AI uni ishlatish ehtimoli kamroq bo'ladi again.

Ishtirok Jazosi

Mavjudlik jarimasi chastota jarimasiga o‘xshaydi, lekin tekis tarzda agar ular sodir bo'lgan yoki bo'lmaganligiga asoslanib tokenlarni jazolaydi, o'rniga proportsional ravishda.

Determinism Eslatma

Harorat va Top-P butunlay nolga o'rnatilganda ham, AI har safar bir xil natijani bermasligi mumkin. Bu AIning "miyasida" amalga oshirilayotgan GPU (grafik protsessor birligi) hisob-kitoblaridagi tasodifiylik tufaylidir.

Xulosa

Xulosa qilib aytganda, temperatura, top p, maksimal uzunlik kabi sozlamalarni o'zlashtirish til modellari bilan ishlashda muhimdir. Ushbu parametrlar modelning chiqishini aniq nazorat qilish imkonini beradi va maxsus vazifalar yoki ilovalarga moslashtirish imkonini beradi. Ular javoblarning tasodifiyligi, javob uzunligi va takrorlanish chastotasi kabi jihatlarni boshqaradi, bularning barchasi AI bilan o'zaro aloqani yaxshilashga hissa qo'shadi.

Qisman jackdickens382 va evintunador tomonidan yozilgan

---

1. Texnikroq so'z "configuration hyperparameters" ↩

2. Shuningdek, Nucleus Sampling nomi bilan ham tanilgan ↩