دانشگاه تهران

دانشکده علوم و فنون نوین

گزارش هفته ششم و هفتم

نام و نام خانوادگی	فاطمه چیت ساز
شماره دانشجویی	830402092
تاریخ ارسال گزارش	6 بهمن 1402

Contents

برسی مقاله week to stronge learner

در این هفته من سعی کردم یکم قشنگ تر مقاله رو بخونم
اولین سوالی که ذهنمو درگیر کرده اینه که اینا دقیقا چطوری میخوان align کنن و اینکه این دادن labelهای مدل ضعیف به قوی دقیقا چه میکنه
وقتی دوباره مقدمه رو خوندم دیدم خیلی تاکید کرده بود روی این قضیه که ما الان reinforcement learningداریم از human feedback و خب چون مدلا دارن میرن سمت superhuman دیگه شاید جواب نباشه خیلی اون یعنی اون مدلا از ما خیلی قوی تر هستند و خب ما میشیم یک سری supervisorکه ضعیفه و خب ممکنه label هاش غلط باشه و حالا باید چه کرد ؟
یعنی این باعث شد که اون دید week to strong generalization بیاد
که ما ی سری مدل ضعیف داریم که میان label میدن در حالی که ما می‌خواهیم مدل قوی اونقدر تحت تاثیر قرار نگیره و وضعیتش نابود شه

مثلا مثال جالبی که خودش زده بود این بود که چار روز دیگه ی خروار کد برای ما میزنه اون مدل قویه و ما ب عنوان مدل ضعیف احتمالا نتونیم label درستی به مدل قویه بدیم

حالا برای این مسئله week to strong اومده مدل قوی رو با label های مدل ضعیف fine-tuning کرده حالا چرا این کار میکنه ؟ چون الان مدل قویه میفهمه مشکلات مدل ضعیف چیه رو چه چیزایی حساسه و دقیقا حرفش چیه دیدش چیه و حتی ی جورایی اورش هم میفهمه (به نظر من این عمق خطرناک بودنه)
ولی خب حرفشون اینه اون مدلی که اون تسک یا حالا کد سخت رو ساخته و درک اینو داشته احتمالا میدونه کدوم کد خوبه کدوم کد بد در واقع مدل ضعیف نیاز نیست بیاد فرایند یادگیری رو انجام بده فقط باید اون برانگیخته کنه اون چیزایی که همین الان مدل قوی می‌دونه

حالا تسک هایی که برای ازمایششون استفاده کردند
بیست و دو تا تسکnlp
تسک شطرنج
تسک reward model gpt هست
که برای مدل قویشون اومدن از gpt چهار استفاده کردند برای ضعیف gpt دو اما یک جا دیدم برای reward modelاومده بود گفته بود ما مدل قویمون 3.5 عه که نفهمیدم چرا ؟

البته یک موضوعی که هست اینه خودشونم فهمیدن شاید الان فقط ی حال proof of concept دارن انجام میدن و توی قدم های اولیه هستند و ما نباید خیلی سخت بگیریم بهشون

خلاصه که یک جورایی برداشت من اینه که preference learning اینا نیازمند ی label خفن از ادمه که چند وقت دیگه این زیاد شدنی نیست و خب باید حل کنیم این مشکل رو

ی بحث جالبی که من زیاد نفهمیدم کرده بود scale oversight بود چیزی که فهمیدم این بود که اینجا برای حالاتی که ادم نمیتونه label خوب بده اومده از یک مدل دیگه خواسته نظرشو راجب خروجی مدل بده یا حتی مسئله رو به چندتا مسئله کوچک تر بشکنه یک همچین چیزی ولی دقیق نفهمیدم الان فرق این با مسئله ما چیه و چرا ما رفتیم سمت generalization و اصلا اینجا مفهوم generalization چیه

بحث جالب بعدی eliciting latent knowledge هستند اینطوری که از مدل نظر صادقانشو میپرسی (یعنی ک چی :/)

میگف ازش تو tampering اینا استفاده میکنن

کلا این مفهوم صداقت و اینا ب نظر فان بود اما دقیق نفهمیدم عملیش چطور میشه

موضوع بعدی بدبختی های قضیه بود

اینکه الان ما هیچ دیدی نسبت به superhuman نداریم و این همه حرف میزنیم ممکنه الکی باشه

بعد اینجا پریدم رفتم قسمت مدلش اینا که قضیش دقیقا چیه

برای تسک nlp گفته بود بیست و دوتا تسک داریم همه چی رو هم باینری کرده و خب برای preprocessingاش هم دیتا ها رو balance کرده و حتی ی سری ها رو حذف کرده

برای مدل اومده قسمت un embedding رو حذف کرده جاش ی مدل خطی گذاشته که وزناش همون وزن un embedding هست و خروجیش دوتا چیزه که خب اینجا فکر کنم صفر و یک میشه من الان این un embedding رو زیاد نمیفهمم قضیش چیه

مدلشم طی دوتا ایپاک با بچ سایز سی و دو ترین میکنه و Early stop هم داره با توجه به label های مدلweek

دیتا ست هم نصف میکنه نصفشو اموزش میده به مدل ضعیف نصفشو تست میکنه به مدل ضعیف لیبل ها رو به عنوان soft label میده مدل قوی

قضیع soft label رو من دقیق نفهمیدم دقیقا چیه

نکته بعدی اینه نمودار سیزده صفحه سی و یک رو من نمیفهمم یعنی الان مدل ضعیف ها بهتر عمل میکنن؟ بعد این چطوری over fitting رو نشون میده

برای مسئله ی شطرنجه اینطوریه که موقعیت مهره ها رو میده میگ بهترین حرکتت چیه
بعد اینا همه temperature صفره یعنی greedy محض دیگه اینکه پنجهاه هزار تا داده رو رندوم انتخاب برای week پنجاه هزار تا رو میده week لیبل میده بعد میده ب strong داده تست هم پنج هزار تا برای boostapingهم پنجاه هزار تا
قضیه اینه توی شکلش میاد نمایش میده مال حالت zero shot رو وبعضی اوقات وقتی week خیلی ضعیفه عملکرد week to strong از zero shot هم کمتر میشه
و خب یک نکته ای که هست قضیه agreement هست
اینکه مدلامون) مدل (student وقتی بزرگ میشه agreement اش کم میشه
این قضیه agreement اینطوریه ک نگا میکنه مدل قوی چقدر داره از مدل ضعیف پیروی میکنه و خب منطقا اگ کپ هم عمل کنن pgr میشه صفر
اما خب ی نمودار باحال داره میگ وقتی مدل دانش آموز قوی میشه وقتی مدل معلم داره دری وری میگ زیاد agreement نمیکنه و اگ aux رو اضافه کنیم دیگ میشه نور علی نور و زیبایی مطلق

برای تسک بعدی یعنی reward model مربوط به gpt اینطوریه که ی دیتا داریم (d,c1,c2,y) اینطوری که d میشه اخرین دیالوگ امده از طرف یوزر وc1 و c2 میشه دوتا جوابی که مدل درست کرده و y اگر یک باشه یعنی ترجیح c2 صفر باشه یعنی یکی دیگ

حالا اینو برای هر دوتا c1 و c2 انجام میدن مدلش اینطوریه باز اومده لایه un embedding رو دست کاری کرده به خطی با یک خروجی میگ این خروجیه logit عه
بعد برای هر دو ی بار مدلو فوروارد و خروجی میشه اختلاف سیگموید این دوتا logit ها

قضیه بعدی generative fine-tuning هست که ما بیایم یک سری چیز رو برجسته کنیم برای مدل ها بدون اینکه بهشون label بدیم در واقع یک فرایند unsupervised هست به صورت fine-tuning با استفاده از تسک های مرتبط به تسک اصلی
مثلا خودش مثالی که زده بود این بود ک ما ی مدل زبانی داریم حالا میخوایم به احساسات اهمیت بیشتری بده پس بهش مثلا کامنت های انلاین ملت رو میدیم ک دستش بیاد احساسات چطوریه
اما من اصلا نمیفهمم چطوری؟
حالا برای همین مسئله reward model ما اینطوریه ک ما مدل قوی و ضعیف رو هر دو رو بعد training میایم با generative fine-tuning اموزش میدیم ک زیبا میکنه انگار مدل ها رو
ساز و کارشم اینطوریه ک مثلا یک پیام یوزر و دوتا جواب براش رو می‌دیم و خب انگار این دوتا جواب بدترین جواب های ممکن بوده ولی ترجیح یوزر رو بهش نمیدیم

قضیه اینه الان مدل یاد میگیره چیو///:

بدترین جواب رو؟

ی حرف دیگه هم زده بود اینکه ممکنه leakage داشته باشیم اینم نمیفهمم چراا

البته این مفهوم leakage رو میاد یکم بعدتر توضیح میده که مثلا سوالای ما ممکنه تو داده های pretrain gptباشه و خب این خوب نی باعث میشه نتایج ما الکی خوب باشه

اما ی حرف عجیبی میزنه اونم اینه ک برای مدل هلی superhuman اینطور نیست اونا اصل داده رو ندارند حالا ممکنه ی چی شبیهشو داشتع باشه و تقلیدی وجود نداره

البته گفته ما برای proof of concept اومدیم ی مدلی رو نشون دادیم ک داده train اش فرق داشتع و از این حرفا

اما میگم جالبیش این بود که میگف ما در superhuman ها وضعمون اینطور نیست ما اونجا self superviseهستیم

ی حرف دیگشم این بود ک prompting اگ استفاده کنیم ممکنه بیشتر تحت تاثیر این قضیه قراره بگیره

حالا سوال پیش میاد prompting چیه

اینطوریه که مثلا ی حال سوالی طور از مدل میپرسه که نظرش راجب فلان چیز چیه بعد جواب مدل رو برسی میکنه دیگ اون دردسر های label بده بره و بیاد رو ندارع

یا حتی میشه few shot قضیه رو دید البته

اما ما چون زیاد اعتماد ب دیتا pretrain مون نیست این زیاد به دردمون نمیخوره فعلا

Dataset	Custom Prompt
CosmosQA	*There is some context, followed by a question about the context, and an answer. Respond with 1 if the answer is correct, and 0 otherwise.*
Deontology	*There is a short scenario, consisting of a situation or request, followed by an excuse. Respond with 1 if the excuse is valid, and with 0 otherwise.*
Justice	*There is a statement. Respond with 1 if it is correct or just and with 0 otherwise.*
QuAIL	*There is a news context, followed by a question about the context, and an answer to the question. Respond with 1 if the answer is correct, and with 0 otherwise.*
SciQ	*There is a science knowledge question, followed by an answer. Respond with 1 if the answer is correct, and with 0 otherwise.*
Social IQa	*There is some context, followed by a social question, followed by an answer. Respond with 1 if the answer is correct, and 0 otherwise.*
Virtue	*There is a short scenario, followed by a judgement of the person involved. Respond with 1 if the judgement is correct, otherwise respond with 0.*

خب حالا یکم راجب auxiliary confidence صحبت کنیم

فرمولش ک بدین صورته

$$L_{\text{conf}}(f) = (1-\alpha) \cdot \text{CE}(f(x), f_w(x)) + \alpha \cdot \text{CE}(f(x), \hat{f}_t(x))$$

قضیه اینه ما ی cross entropy داریم ک میاد توزیع بین مدل ضعیف و مدل قوی رو برسی میکنه و این میشه loss اش حالا بحث بعدی اینه ما برای اینکه قضیه رو متعادل تر کنیم ی hardened strong model predictionمیسازیم ک خروجیش یا یکه یا صفر و اینطوریه ک اگ f(x) بیشتر از یک thresholdباشه یک وگرنه صفر برمیگردونه و خب میگه این کمک میکنه به balance قضیه

در قسمت بعد اومده ی سری روش هایی ک خیلی جواب نداده اما شاید جواب بده بعدا رو معرفی میکنه اولیش اینه ی threshold بدیم بگیم اینا رو مدل ضعیف خوب فهمیده حالا بیا با همین اینایی ک خوب فهمیده مدل قوی رو اموزش بده ولی این مال وقتیه که مدل ضعیف خیلی وضعیت داغونی داشته باشه و وقتی مدل ضعیف بدک نیست اونقد تاثیری ندارع

روش بعدی productive confidence loss هست که توش میخوایم cross entropy ما له حاصل ضرب دوتا احتمال های مدل ضعیف و قوی برسه که خب اینم بعضی جاها خوب بود بعضی جاها زیاد خوب نبود

موضوع بعدی اینه که بیایم یک مدل خطی از مدل قوی بسازیم بعد تمام لایه ها رو با مقادیر اون-fine tuningکنیم

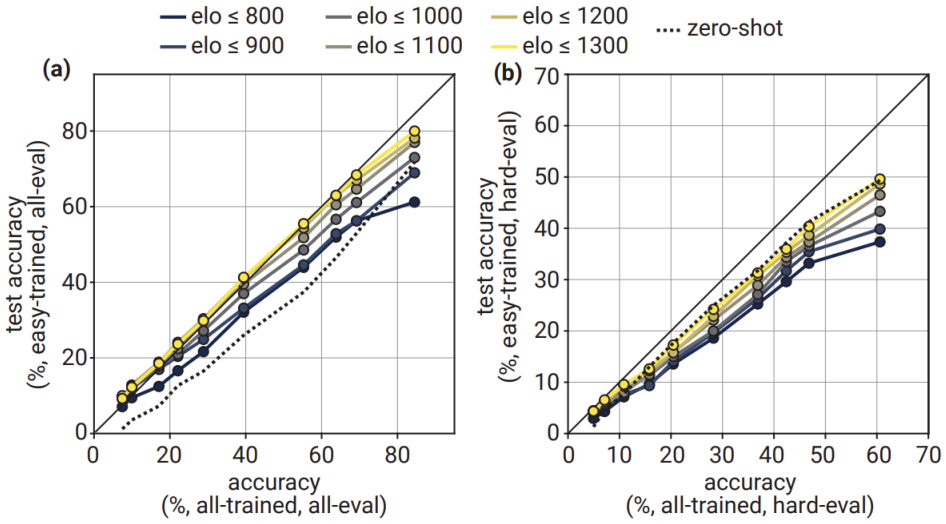
روش های بعدی weight regularization هست مثل lora اینطوری که مثلا یک مدل خطی روی پارامتر ها داریم و ی جورایی همه وزن ها رو نگهداری نمیکنیم انگار به صورت محدودتری میایم-fine tuningمیکنیم

روش های data augmentation اینطوریه که مثلا یک مدل قوی داریم که میاد داده ها رو تقویت میکنه بعد خروجی این مدل قوی و مدل اصلی باهم مقایسه میشه و خب axillary loss رو تفاوت این دوتا ایجاد میکنه

مسئله بعدی easy to hard هست که جالبه مثلا اگر ما مسئله رو با سوالای ساده برای مدل ضعیف fine-tuningکنیم عملکردش از zero shot بهتره

اما برای مسائل سخت اگر این کارو بکنیم اونقد کارمون با zero shot فرقی نداره

در واقع جالبه وقتی مدل در سوالای سخت حتی هر چی مدل ضعیف یادگرفته هم میگیره بازم اونقد از zero shotبهتر نی



مسئله بعدی اینه ک ما داریم دقیقا چی کار میکنیم/:

یعنی الان چطوری میتونیم یک موضوعی را برسی کنیم وقتی مشکل رو الان نداریم

خب باید ی سری چیز رو رعایت کنیم

مثلا فرضیاتمون مشخص باشه و سعی کنیم برای فرض های مختلف راه حل ارائه بدیم

بدونیم برای هر چیز چه گیر و گور هایی داریم

راه حل هامون scalable باشه

در واقع یک دلیلی هم ک رفتن سراغ fine-tuning نرفتن سراغ few shot همینه

توی fine tuning اینطوریه اونی که مدل الان میدونه رو میخوایم دقتش رو ببریم بالا بدون اینکه بدونیم اینی که میدونه رو از کجا یاد گرفته

اما تو few shot اینطوری نی میگ با اطلاعات کم بیا یک توزیع گسترده از داده ها رو بساز

حالا بعدش یکم راجب این میگ ک مسیر چیه

یک اینکه ما باید بیایم align کنیم

حالا راه حل های موجود

RLHF
Scalable oversight
Constitution ai
Adversarial training

ک ما اومدیم گیر دادیم به RLHF اینجا

بعد حالا که align کردیم برسی کنیم ایا درست align کردیم یا نه ؟

روش های موجود

Red teaming

Interpretability

بعد همین فرایند رو هی انجام بدیم و بزرگ و بزرگ تر کنیم باbootstrapping

ی سری چیزا هم برای ما مهمه مثل

امنیت

صداقت

رعایت پروتوکل های انسانی

امنیت کد

که توقع میره که اون مدل های قوی خیلی درک بالایی از اینا داشته باشند

در واقع به صورت ایده ال اونا انقد این مفاهیم رو میفهمند که میشه تبدیلش کرد به یکreward model

ولی اگر اینطور نشد میشه از یک مفهومی به نام oracle استفاده کرد

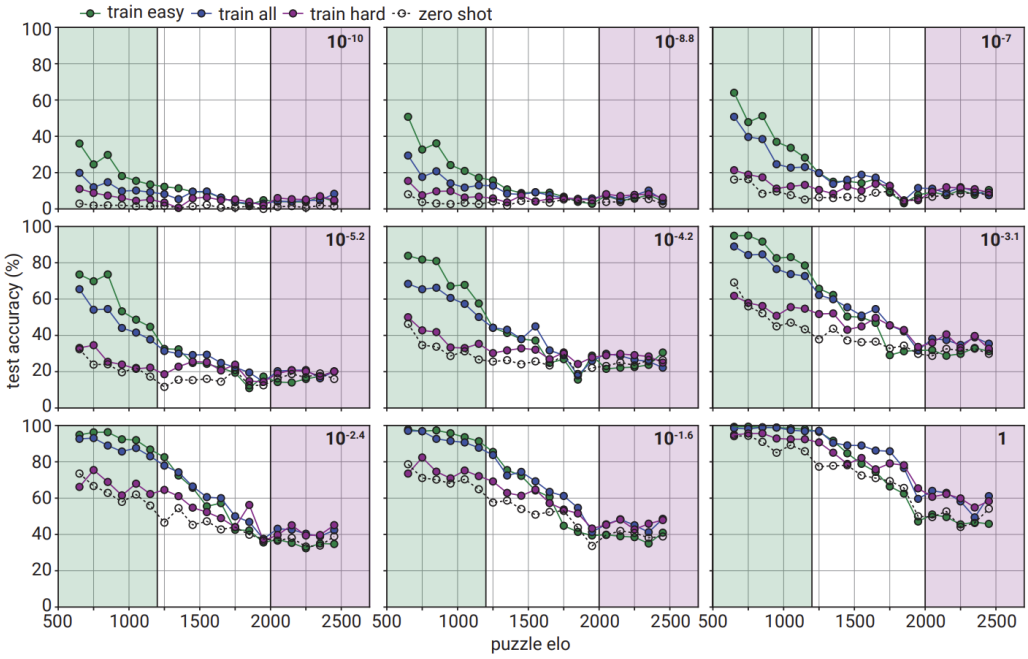
در واقع حرفشون بوی این رو میداد که اگر انقدر superhuman چرا از خودش نپرسیم که وضعیت چطوره و چطوری میشه مشکل را حل کرد

مطلب بعدی easy to hard generalization هست اینطوری که میگ بعضی خطا ها ممکنه برای ما خطرناک باشه مث خطاهای سیستمی

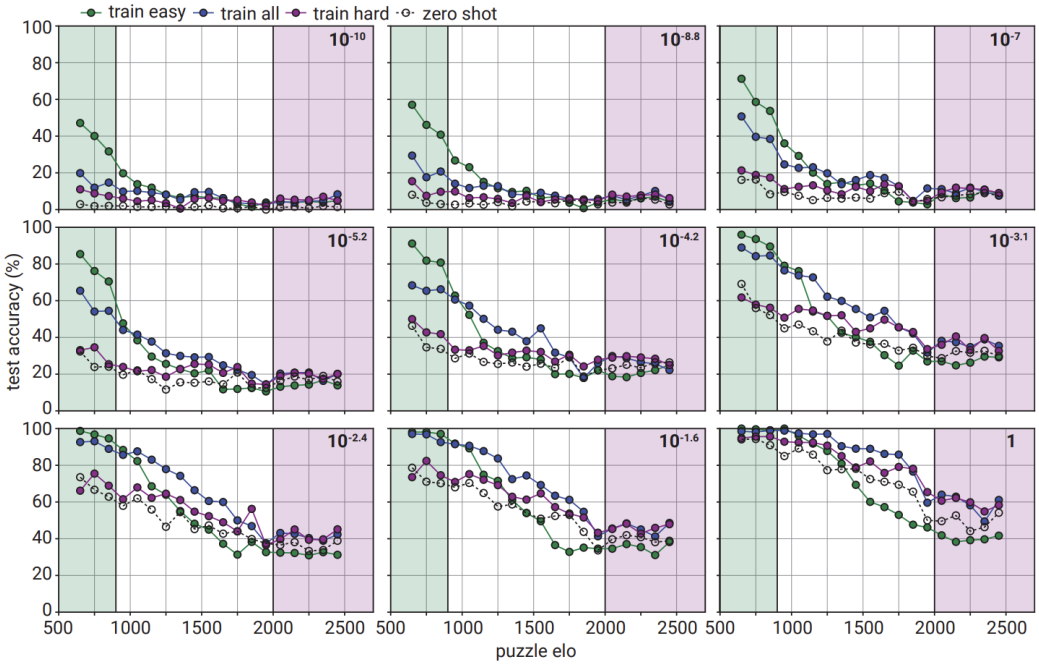
که نمونه ای از اینا اینطوریه که مدل ما برای مسئله های آسون عالی عمل می‌کنه برای مسئله های سخت افتضاح

خلاصه برای برسی این مثلا معیار سختی elo رو برای مسئله chess معرفی میکنه بدین صورت که سخت میشه elo دو هزار به بالا

ولی خب ی مسئله هست اونم اینه این معیار سختی از نظر ادمه نه مدل و ممکنه برای مدل فرق بکنه (ولی خب سوالی که من دارم اینه الان week ادمه پس خوبه که سختی بر اساس ادمه؟ و یک ایده جالبه به نظرم برسی مسئله های سختی که برای ادم سخت برای مدل اسون و خب ی جورایی حالا مدل باهوش تر از ما میشه نه؟؟)



(a) Easy cutoff: Elo $\leq$ 1200



(b) Easy cutoff: Elo $\leq$ 900

خلاصه که برای بازه های مختلف elo و مدل های قوی مختلف (البته من نمیدونم مدل ضعیفه یا قویه ؟) بعد هیچی برسیات میکنه و خب zero shot هم نشون داده و یک چیز فان دیگم که متغیر کرده اینه که ما روی داده های اسون train شیم فقط یا روی فقط سختا یا روی همه

اره خلاصه که وقتی سوالا سخت میشن منطقا accuracy گند میخوره(((:

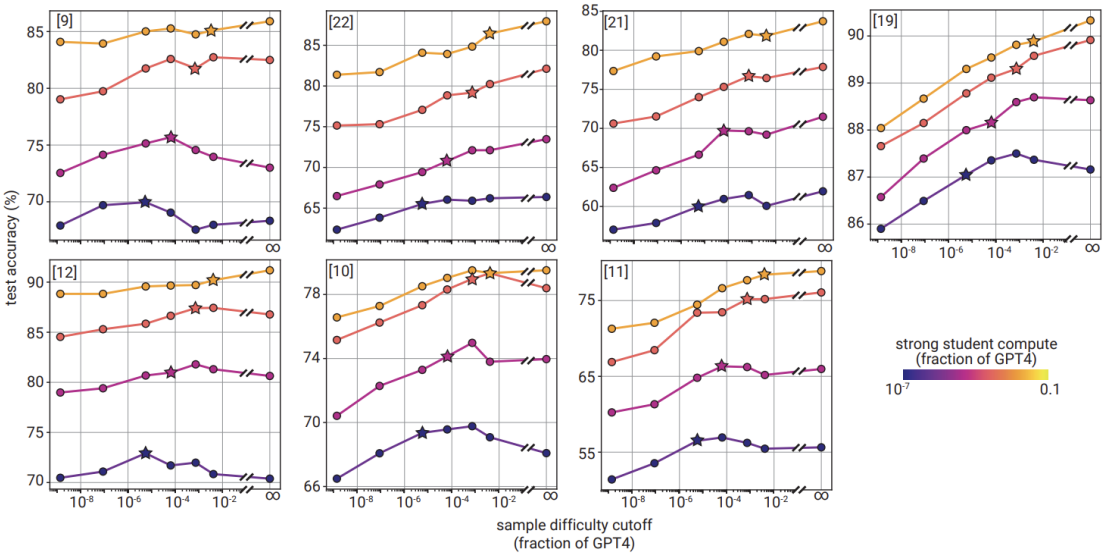
ولی نکته ای که باحاله اینه دقت در سوالای آسون برای مدل هایی که train easy شدن بیشتره(lol)

ولی وقتی سخت میشه قضیه معمولا train hard بهتره البته همشون معمولا گند زدن و حتی در حد zero shotهستند

یک کار دیگه هم اینجا میکنه اونم اینه که مسئله رو اصلا بیایم بدون مشکل ارور های week label در نظر بگیریم و خب فرض کنیم مدل ها دسترسی دارند به Grand Truth حالا برای مسئله های nlp ما معیار سختی را اندازه کوچکترین مدلی بگیریم که سوال رو برای یک دیتا پوینت همیشه درست جواب میده

حالا که میایم مدل ها رو اموزش میدیم با دسته های مختلف سختی خب به طور کلی هر چی با سخت تر ها اموزش بدیم accuracy بهتر میشه اما بعضی جاها برای مدل ضعیفا اینطوریه ک از ی جایی به بعد که مسئله سخت میشه مدل دیگ کشش نداره با توجه به سایزش و خب حتی وقتی با چیزای خفن تر trainمیکنیمش دقتش میاد پایین

(درست متوجه شدم ؟؟؟ train hard در مدل های ضعیف باعث پایین آمدن accuracy میشه؟؟؟؟ خب این چرا باید به درد ما بخوره؟)



موضوع بعدی که هست اینه که ما برای همه چی عین هلو معیار مشخص سختی نداریم و خب سخته پیدا کردنlabel

راه حل اینه مثلا از gpt چهار بپرسیم بین فلان سوال یک و سوال دو کدوم سخت تر هست برای انسان منطقا کاملا شبیه انسان نیست ولی خب

نکته جالب اینه درجه سختی بالایی میده برای prompt های بزرگ (در حالی که نویسنده نظرش این بوده برای prompt های بزرگ ممکنه مدل کوچک بهتر بفهمه حتی چون context بیشتری داره که من زیاد موافق نیستم)

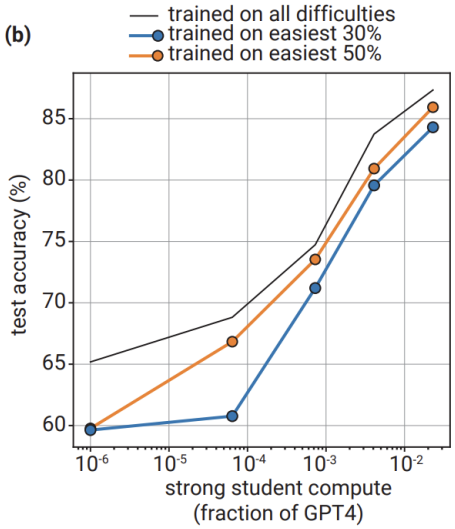
(الان من یکم نفهمیدم ما درجه سختی ی چیز از نظر انسان برامون مهمه یا مدل ؟؟؟)

(های اصلی label کردن با سخت تست (البته train خلاصه اومدن مدل رو با داده های ساده

دیدن ماشالله چه شیبی

ما نویزیه دلیلش همونه elo بعد نظرشون این بود ک چون این داده های

ولی میگن این نویزه کمک کنندست من نمی‌فهمم چرا!؟ چون شیب یادگیری و دقت زیاد بود؟؟؟



قضیه بعدی اینه اومدن ی سری مدل مال image رو برسی کردند مدل ضعیفشون Alex net هست دادشون image net مدل های قویشون ی حال self supervise دارند پسسس leakage نداریم res net 50و vit -b مدل های قوی هستند

حالا یک مار باحالی که کردند اینه که از linear prob استفاده کردند اینطوری ک وزن ها همه فریز

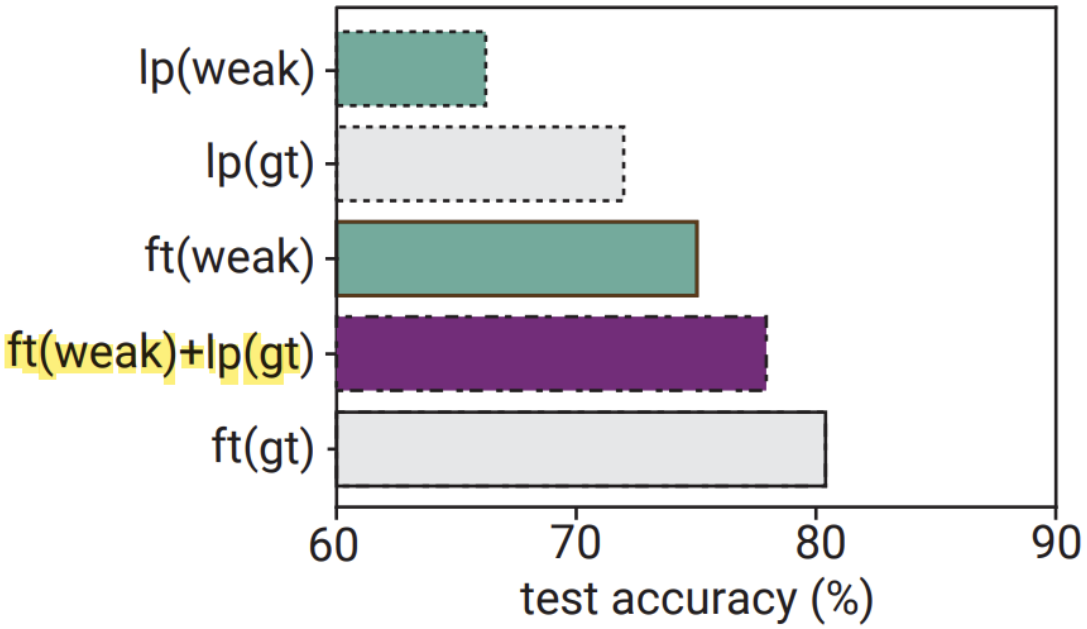
میشن و یک لایه خطی اموزش داده میشه یک بار با لیبل های اصلی و بک بار با لیبل های مدل ضعیف و خب این مدل خطی رو ما tarin میکنیم و وزن هاشو پیدا میکنیم

کلا من یکم نمیفهمم الان این مدل خطیه میره کجا؟ سر مدل؟؟ چرا؟

ولی خب میگف خیلی ب نتایج fine tuning سریع تر میرسه و این حرفا

قضیه بعدی ترکیب این با fine tuning هست

یعنی میشه مدل ضعیف خطی کرد یا مدل قوی را خطی کرد یا مدل را با لیبل ضعیف fine tuning کرد یا مدل قوی را ک ft کردیم حالا با لیبل های اصلی lpاکنیم یا کلا ft روی لیبل اصلی ک خب نتایج میشه این ریختی



نکته ای که هست اینه هنوز ft روی لیبل اصلی شاخ ترع

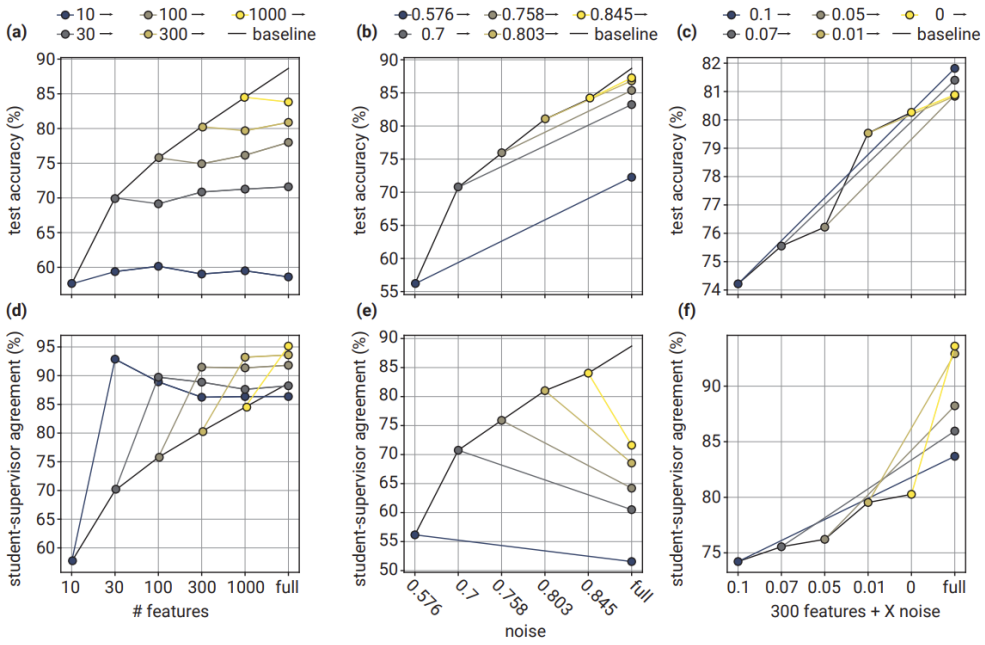
نکته بعدی اینه ft(week) با lp خیلی باحاله

نکته بعدیش اینه وقتی مدل های ما طی یک سری فرایند ft میشن ولی label های ft اونقدر خفن نی باعث میشه مدل خطی درنشون خوب عمل کنه

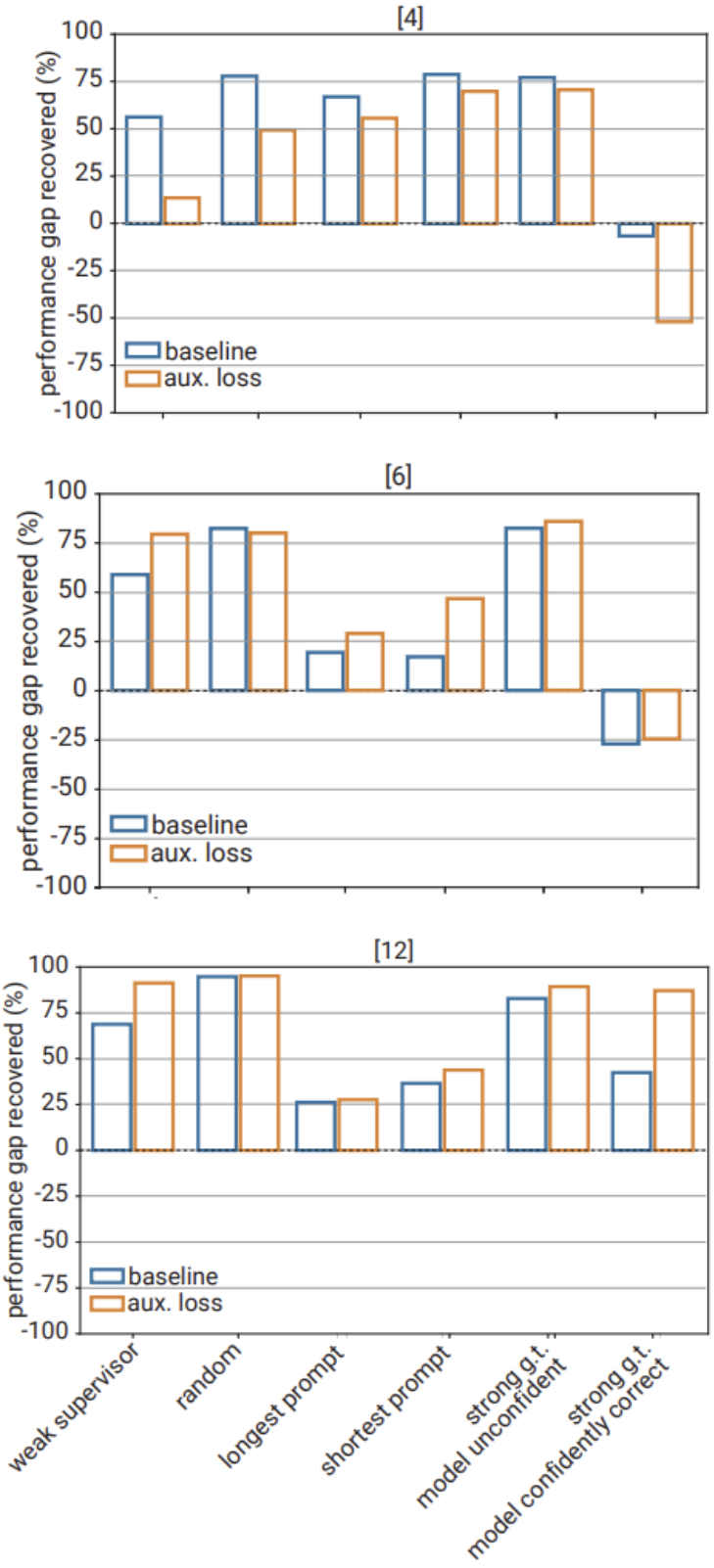
در قسمت بعد اومده یکم راجب ارور های مختلف مدل ضعیف و تاثیر اونها بر روی مدل قوی صحبت کرده

برای این کار اومده خیلی حالت مصنوعی طور یک سری ازمایشات کرده اینطوری که مثلا مدل های خطیش فقط بخشی از فیچر ها رو ببینند در این حالت ارور مربوط به لیبل های مدل ضعیف چه تاثیری میذاره روی مدل قوی --- > مشاهده میشه که خیلی وضع مدل قوی بهتر از مدل ضعیف نیست و aggrement زیادی داره

اگر مثلا مدل ضعیف رندوم طور یک سری لیبل ها رو غلط بگه چی --> **اون وقت مدل قوی زیبا عمل میکنه و با مدل ضعیف زیاد موافق نیست و** accuracy **ها هم خوبن**

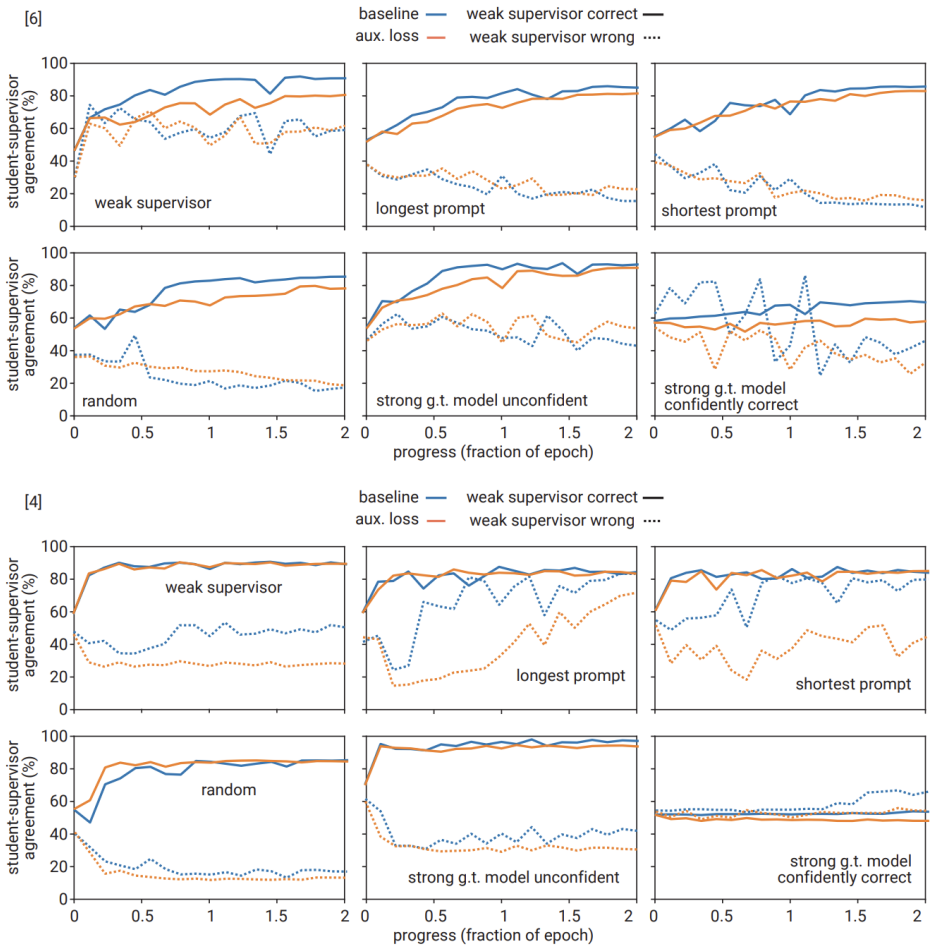


ارور های مختلف و تاثیر اونا


[4]
performance gap recovered (%)
100
75
50
25
0
-25
-50
-75
-100
baseline
aux. loss
[6]
performance gap recovered (%)
100
75
50
25
0
-25
-50
-75
-100
baseline
aux. loss
[12]
performance gap recovered (%)
100
75
50
25
0
-25
-50
-75
-100
baseline
aux. loss
weak supervisor
random
longest prompt
shortest prompt
strong g.t. model unconfident
strong g.t. model confidently correct

وقتی به صورت ی حال بی اهمیتی از مدل میپرسیم لیبل اینه نظر تو چیه

(من خیلی نفهمیدم الان منظورش از این کار چیه گفته بود چون مدل های super human در تقلید از ادم خیلی خوبن پس ممکنه بد شه قضیه)



چندتا مسئله

یک استاد احمدی گفت nlp ارائه میشه یک نیشخندی هم بهم زد

دو اینکه پنج شنبه ها میخوام پاشم بیام دانشگاه

سه اینکه زده به سرم کلا کار رو ول کن مثل اوسکولا برم کاراموز هوش شم