検索アイテムの付加価値推定と LLM での応用

橋本 和真 / Kazuma HASHIMOTO

Google

NLP コロキウム @ 2024 年 10 月 30 日 (水)

公開用・改訂版 (2024 年 11 月 1 日)

自己紹介

- 2012～2018: 東京大学
  - 卒論～博士 (指導教員: 鶴岡 慶雅 先生)
- 2018～2021: Salesforce
  - AI Research
- 2021～現在: Google
  - Research (2021～2024 前半)
  - DeepMind (2024 後半～)

+ 技術系・研究の世界に興味のある学生と研究を行う非営利組織のボランティア活動

- 個人ページ:
https://hassygo.github.io
- メール:
kazumah@google.com
- LinkedIn:
Kazuma Hashimoto

アメリカのベイエリアが拠点なので、
お越しの際はお気軽に声をかけてください。

今回の発表の流れ

1. 検索ベースの NLP モデルと応用に関する話
2. 企業の LLM 研究の中心地で働くことに関する話

検索ベースの NLP モデル

- 検索 (search, retrieval) を利用した NLP タスク・モデル
  - QA: query -> **document(s)** -> answer
  - MT: source text -> **translation memory** -> translation
  - …

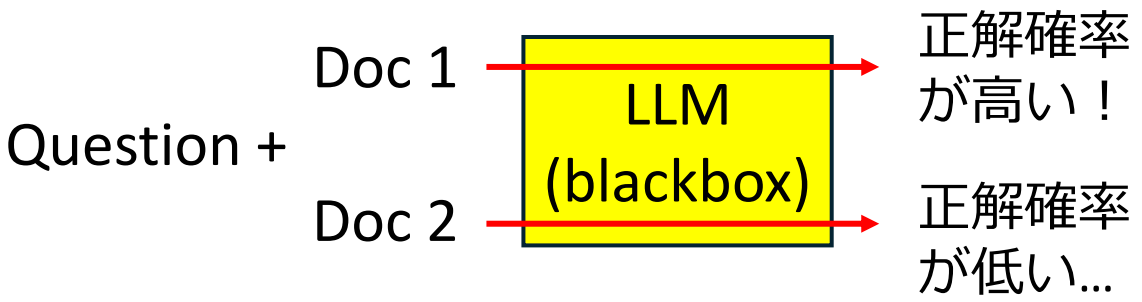
- RAG (retrieval-augmented generation) という形で更に普及
  - 昨今の LLM (large language model) と相性が良く、様々な問題において効果的に使用可能

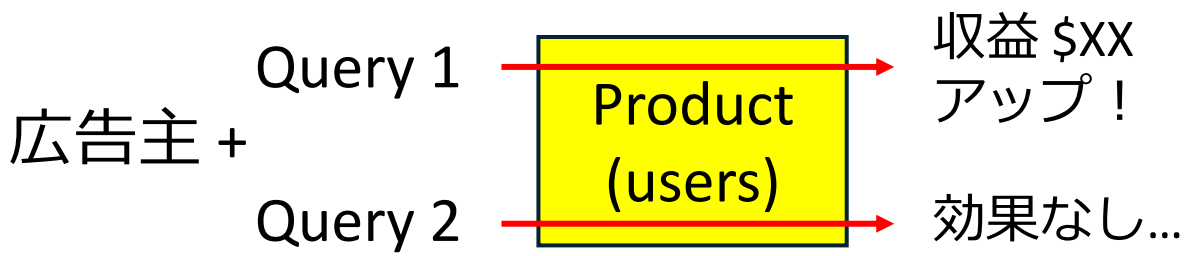
検索クエリ自体や結果の**妥当性 (価値)** はどのように測られるのか？

検索クエリ・結果の価値の推定

- (半) 人手のアノテーション
  - 研究用のデータセットでは各 query に対して有用なデータのアノテーションが付与されることがある
  - 例: question + **gold docs** -> retrieval/reranking モデルの学習

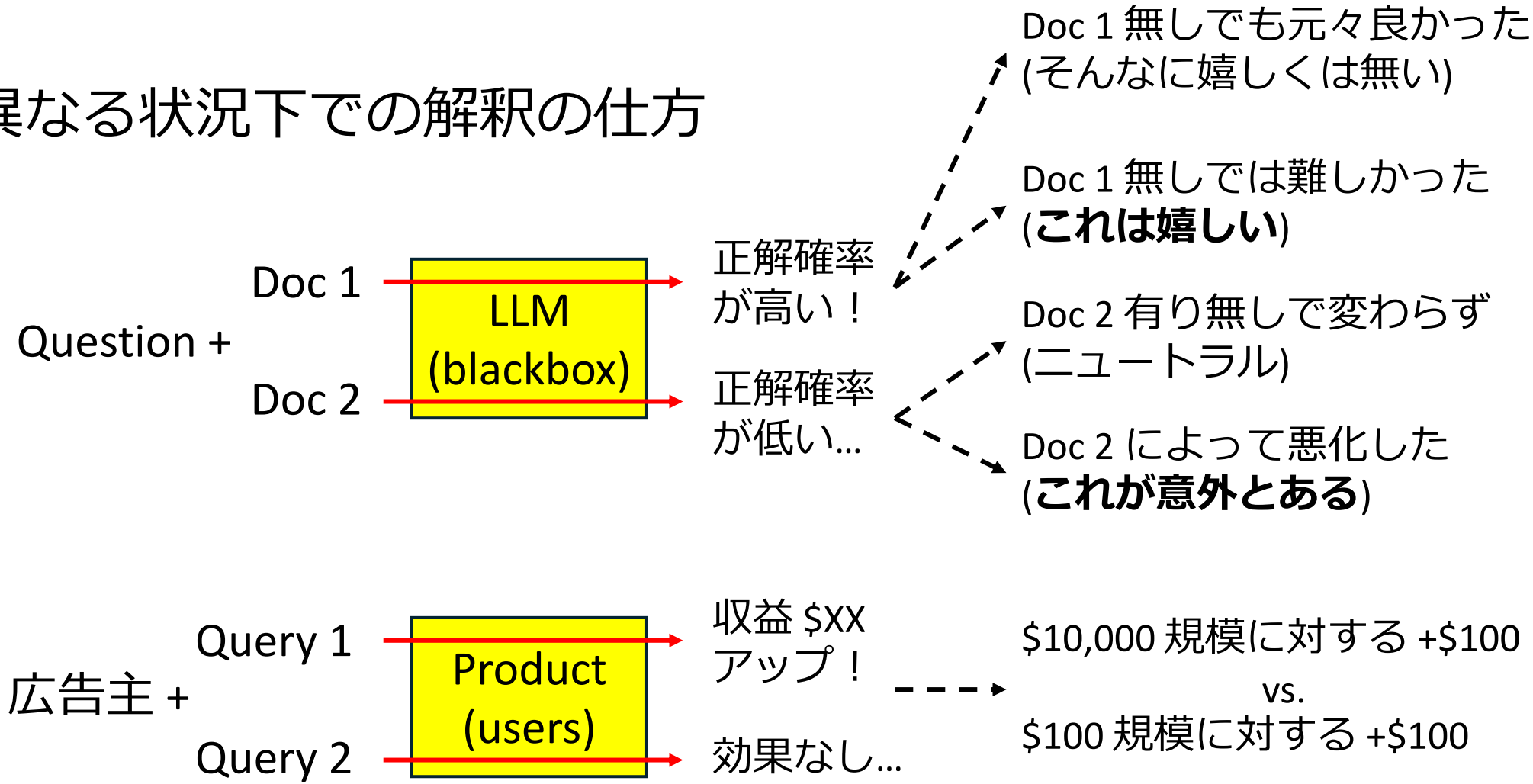
- 最終ゴールの達成度・報酬によるアノテーション (**今回の話題**)





ゴール達成度・報酬の解釈

➢異なる状況下での解釈の仕方



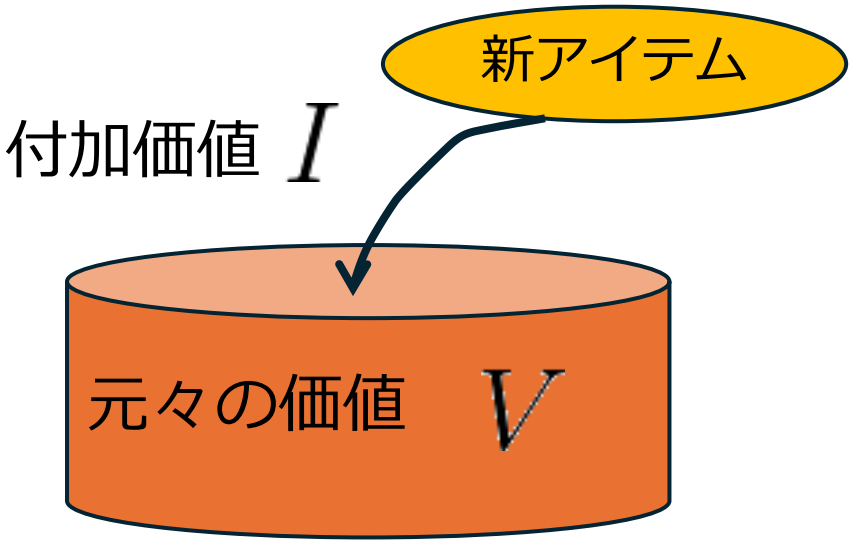
付加価値 (Incremental Utility)

- ゴールの達成度、スコア、報酬などの絶対値ではなく、新しいアイテムによってもたらされた**付加価値の割合**を計算

$$\frac{(V+I)-V}{V+I}$$

問題によっては
- スムージング
- リスケーリング

等を行う。



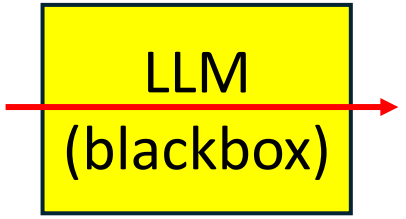
In-Context Learning での話

- ICL (in-context learning)
  - LLM に**入出力の事例**を与えることで新しい入力に対処
  - 手続き的には RAG に近い (有用な事例を見つけたい)
  - 各事例の付加価値を計算したい

入出力の事例 (demonstrations)

Example input 1 -> output 1
Example input 2 -> output 2
…
Example input *K* -> output *K*

+ test input

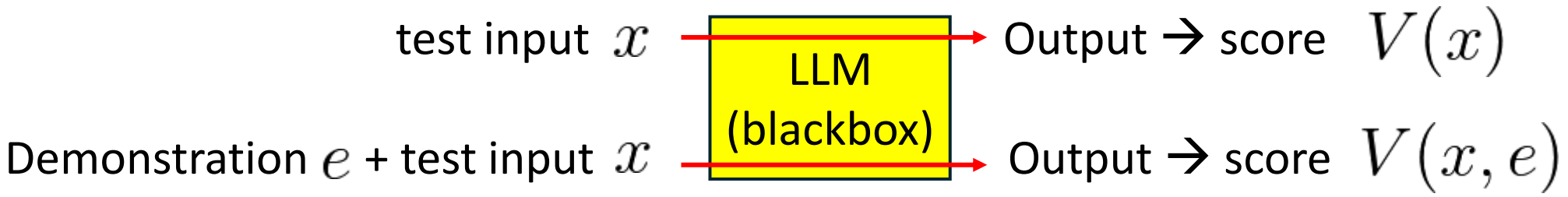


Output

どの事例の価値が高いのか？

入出力事例の付加価値

➢各入出力事例を足した場合の LLM の出力への影響を計算
- 入出力**事例間の関係**が考慮されていないのでベストではない



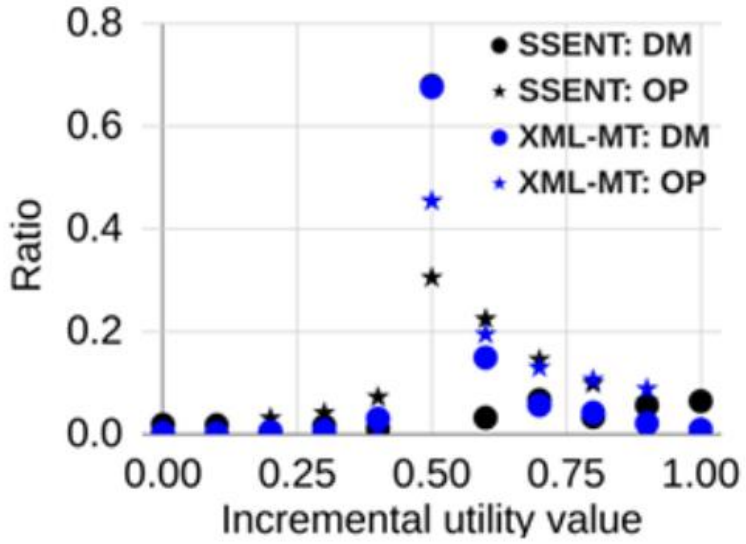
$$\frac{V(x,e) - V(x)}{\max(V(x,e), V(x))^{\ell}}$$

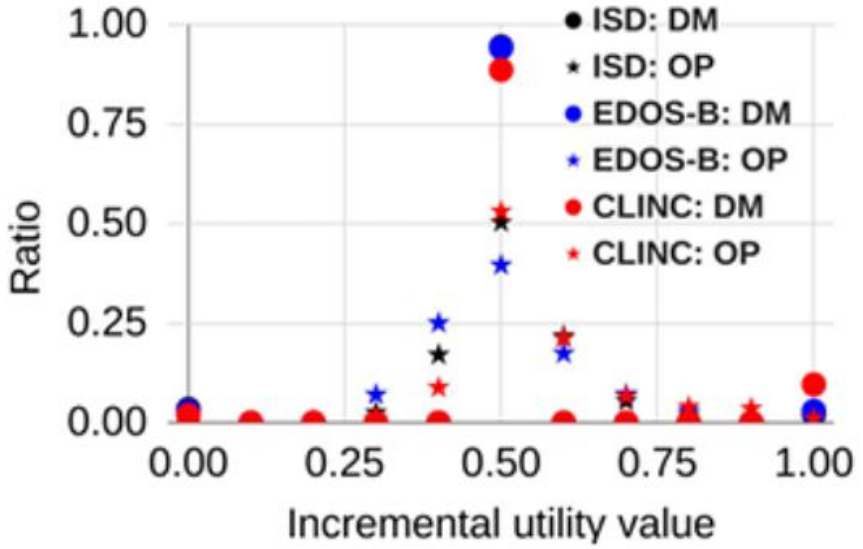
値域は [-1.0, 1.0]
→
値域が [0.0, 1.0] になるように線形に変換

(分母の変形は、**値の対称性**とリスケーリングのため)

付加価値推定の実例

➢様々なデータセットで入出力の価値推定をプロット
- DM: V = タスク特有の評価指標 (F1, BLEU, etc.)
- OP: V = 正解例を出力する場合の確率





学習データでこの分析を行い、
その結果を元に**未知のデータに関しても価値の高い入出力事例を選択**できるように学習 (NAACL 2024)

マルチタスク・多言語への対応

➢新しい問題に対して毎回この分析＋モデル学習は高コスト

➢Demonstration 選択モデルの汎化性能アップの検討
- Reranking model → dense retrieval model (スケーラビリティ)
- 1 モデルを **80 以上のデータセット** (タスク) で学習
- 新規タスクに対応する際のチャレンジ
- マルチ・クロスリンガル性
  - 最新の Google Translate を利用して **230 の言語へ対応**すべく ICL のデータ拡張

様々なタスクへの対応の難しさ

- 80 の多様なデータセットを収集
  - 要約、翻訳、NLI、・・・
  - マルチリンガル

- Retrieval モデルを使う際の前提
  - タスクの instruction が簡素に書ける
  - 入出力の形式がはっきりしている

- 最近の LLM が扱う問題はもっと複雑
  - きれいな形式のデータセットで学習したモデルが通用しない (**現在の状況**)

No.	Name	Type	Languages	Source	Scoring	$\|Q_j\|$	$\|C_j\|$
01	WMT14 cs→fr (Bojar et al., 2014)	Machine translation	cs, fr	Link	GLEU	100,000	30,059,732
02	WMT14 fr→cs (Bojar et al., 2014)	Machine translation	cs, fr	Link	GLEU	100,000	30,059,732
03	WMT16 cs→de (Bojar et al., 2016)	Machine translation	de, cs	Link	GLEU	60,000	4,143,251
04	WMT16 de→cs (Bojar et al., 2016)	Machine translation	de, cs	Link	GLEU	60,000	4,143,251
05	WMT16 cs→ru (Bojar et al., 2016)	Machine translation	cs, ru	Link	GLEU	30,000	2,296,592
06	WMT16 ru→cs (Bojar et al., 2016)	Machine translation	cs, ru	Link	GLEU	30,000	2,296,592
07	ANLI r1 (Nie et al., 2020)	Natural language inference	en [+MT]	Link	Probability	8,473	8,473
08	ANLI r2 (Nie et al., 2020)	Natural language inference	en	Link	Probability	22,730	22,730
09	ANLI r3 (Nie et al., 2020)	Natural language inference	en	Link	Probability	30,000	70,459
10	QNLI (Rajpurkar et al., 2018)	Natural language inference	en	Link	Probability	30,000	74,543
11	MNLI (Williams et al., 2018)	Natural language inference	en	Link	Probability	30,000	100,000
12	WNLI (Levesque et al., 2012a)	Natural language inference	en	Link	Probability	317	318
13	MRPC (Dolan and Brockett, 2005)	Paraphrase identification	en	Link	Probability	200	3,268
14	PAWS (Zhang et al., 2019)	Paraphrase identification	en	Link	Probability	30,000	19,401
15	Tatoeba (Artetxe and Schwenk, 2019)	Translation identification	sqi, fry, kur, tur, ...	Link	Probability	30,000	177,554
16	IMDB (Maas et al., 2011)	Sentiment classification	en	Link	Probability	12,400	12,400
17	SST2 (Socher et al., 2013)	Sentiment classification	en	Link	Probability	30,000	37,149
18	Yelp (Fast AI)	Sentiment classification	en	Link	Probability	30,000	100,000
19	Tweet Sentiment Extraction (Kaggle)	Sentiment classification	en [+MT]	Link	Probability	10,000	17,281
20	AfriSenti (Muhammad et al., 2023a)	Sentiment classification	amh, hau, ibo, ...	Link	Probability	30,000	33,685
21	TweetEval-emoji (Barbieri et al., 2018)	Emoji classification	en	Link	Probability	20,000	25,000
22	TweetEval-emotion (Mohammad et al., 2018)	Emotion classification	en	Link	Probability	1,600	1,657
23	DialogEmotion (Kumar et al., 2024)	Multi-speaker emotion classification	en, hi	Link	F1	700	799
24	Massive-intent (FitzGerald et al., 2022)	Dialog intent classification	af, am, ar, az, ...	Link	Probability	30,000	100,000
25	MTOP-domain (Li et al., 2021)	Dialog domain classification	de, en, es, fr, ...	Link	Probability	30,000	43,928
26	MTOP-intent (Li et al., 2021)	Dialog intent classification	de, en, es, fr, ...	Link	Probability	30,000	43,928
27	ATIS-intent (Price, 1990)	Multi-label dialog intent classification	en	Link	F1	2,000	2,189
28	E2ENLG-reversed (Dušek et al., 2019)	Semantic parsing (text to dict)	en	Link	F1	16,662	16663
29	WikiSQL (Zhong et al., 2017)	Semantic parsing (text/table to SQL)	en	Link	GLEU	20,000	36,355
30	BC5CDR (Li et al., 2016)	Named entity recognition (biomedical)	en	Link	F1	2,000	2,560
31	BioNLP13PC (Ohta et al., 2013)	Named entity recognition (biomedical)	en	Link	F1	1,000	1,499
32	JNLPBA (Huang et al., 2020)	Named entity recognition (biomedical)	en	Link	F1	9,000	9,346
33	MultiCoNER2 (Fetahu et al., 2023)	Named entity recognition	de, fa, fr, ...	Link	F1	30,000	140,824
34	CoNLL2003 (Tjong Kim Sang and De Meulder, 2003)	Named entity recognition	en	Link	F1	7,000	7,041
35	MTOP-slot (Li et al., 2021)	Dialog slot labeling	en, fr, hi	Link	F1	19,000	19,811
36	SNIPS-slot (Coucke et al., 2018)	Dialog slot labeling	en	Link	F1	6,000	7,084
37	ATIS-slot (Price, 1990)	Dialog slot labeling	en	Link	F1	2,000	2,478
38	SemRel (Hendrickx et al., 2010)	Relation classification (nominals)	en [+MT]	Link	Probability	3,800	4,000
39	DDI13 (Herrero-Zazo et al., 2013)	Relation classification (drugs)	en	Link	Probability	8,000	10,779
40	ChemProt (Islamaj Doğan et al., 2019)	Relation classification (chemical and protein)	en	Link	Probability	9,000	10,460
41	WordSeg (Bahlo et al., 2020)	Word segmentation	en	Link	GLEU	30,000	100,000
42	FixPunct (Bahlo et al., 2020)	Punctuation fix	en	Link	GLEU	30,000	100,000
43	CoLA (Warstadt et al., 2019)	Linguistic acceptability judgment	en	Link	Probability	4,175	4,176
44	CoNLL2000 (Tjong Kim Sang and Buchholz, 2000)	Syntactic phrase chunking	en	Link	F1	4,000	4,936
45	Pronoun (Rahman and Ng, 2012)	Coreference resolution	en	Link	Probability	561	561
46	WSC (Levesque et al., 2012b)	Coreference resolution	en	Link	Probability	252	252
47	WinoGrande (Sakaguchi et al., 2019)	Sentence completion	en	Link	Probability	20,099	20,099
48	WiC (Pilehvar and Camacho-Collados, 2019)	Word sense disambiguation	en	Link	Probability	2,614	2,614
49	Python (Lu et al., 2021)	Code summarization	en	Link	GLEU	30,000	100,000
50	Java (Lu et al., 2021)	Code summarization	en	Link	GLEU	30,000	100,000
51	Go (Lu et al., 2021)	Code summarization	en	Link	GLEU	30,000	100,000
52	PHP (Lu et al., 2021)	Code summarization	en	Link	GLEU	30,000	100,000
53	Gigaword (Napoles et al., 2012)	Text summarization	en	Link	GLEU	30,000	100,000
54	SAMSum (Gliwa et al., 2019)	Dialog summarization	en	Link	GLEU	7,366	7,366
55	iDebate (Wang and Ling, 2016)	Debate summarization	en [+MT]	Link	GLEU	859	800
56	MultiHateClock (Röttger et al., 2022)	Hate speech detection/classification	en, fr, hi, it, ...	Link	Probability	20,055	20,055
57	Toxic (Muennighoff et al., 2023)	Toxic text detection	en	Link	Probability	24,900	24,900
58	Counfact (O'Neill et al., 2021)	Counterfactual review detection	de, en, ja	Link	Probability	7,500	7,718
59	Irony (Van Hee et al., 2018)	Irony detection	en	Link	Probability	1,400	1,462
60	Offensive (Zampieri et al., 2019)	Offensive text detection	en	Link	Probability	5,000	6,916
61	Sarcasm (Abu Farha et al., 2022)	Sarcasm detection	ar, en	Link	Probability	2,500	3,414
62	SQuAD2 (Rajpurkar et al., 2018)	Reading comprehension	en	Link	GLEU	30,000	100,119
63	BoolQ (Clark et al., 2019)	Reading comprehension	en [+MT]	Link	Probability	4,613	4,614
64	DROP (Dua et al., 2019)	Reading comprehension (numerical)	en	Link	Probability	29,635	46,621
65	OpenbookQA (Mihaylov et al., 2018)	Reading comprehension	en	Link	Probability	2,478	2,478
66	Cosmos (Huang et al., 2019)	Reading comprehension (common sense)	en	Link	Probability	12,531	12,531
67	SciDocs (Cohan et al., 2020)	Relevance, re-ranking	en	Link	Probability	30,000	99,159
68	HotpotQA (Yang et al., 2018)	Relevance, re-ranking	en	Link	F1	30,000	60,447
69	AI2 ARC-easy (Clark et al., 2018)	Closed-book question answering	en	Link	Probability	1,025	1,026
70	AI2 ARC-challenge (Clark et al., 2018)	Closed-book question answering	en	Link	Probability	459	460
71	TriviaQA (Joshi et al., 2017)	Closed-book question answering	en	Link	Probability	30,000	108,184
72	Math (Saxton et al., 2019)	Math question answering	en	Link	Probability	30,000	100,000
73	CommonGen (Lin et al., 2020)	Constrained text generation (common sense)	en	Link	GLEU	30,000	37,189
74	SNLI-en (Bowman et al., 2015)	Constrained text generation (entailment)	en	Link	GLEU	10,112	33,106
75	PIQA-ques (Bisk et al., 2019)	Question/query generation	en [+MT]	Link	GLEU	7,956	7,957
76	arXiv (Muennighoff et al., 2023)	Multi-label topic/category classification	en	Link	F1	30,000	69,113
77	medRxiv (Muennighoff et al., 2023)	Topic/category classification	en	Link	Probability	5,000	16,229
78	DBpedia (Lehmann et al., 2014)	Topic/category classification	en [+MT]	Link	Probability	5,000	5,000
79	Yahoo (Zhang et al., 2015)	Topic/category classification	en	Link	Probability	14,575	14,575
80	AG news (Zhang et al., 2015)	Topic/category classification	en	Link	Probability	30,000	89,800
81	TREC (Li and Roth, 2002)	Topic/category classification	en [+MT]	Link	Probability	2,626	2,626

以上を基にした最近の興味

- 従来の事前学習モデル (mT5 等) の retriever に限界を感じる
  - 「要約」「翻訳」などとしたざっくりしたタスク instruction ではなく、「〇〇に特に気を付けた要約/翻訳」といった **ニュアンスを加味したうえで価値のある事例**を選択可能か？
    - より最近の product レベルの LLM を用いることで可能かどうか

- これまでの研究の LLM 学習自体への還元
  - LLM を blackbox として feedback を得るために使うのではなく、それ自体に価値推定などの能力を明示的に学習させる

後半の話題

- LLM 研究の中心により近づくことで研究の仕事が変化
  - Google Research 時代:
  - 研究 <--> Product の繋がりを意識しつつも Publication は推奨されていた

- 現在は研究と競争意識の高い Product (LLM) との距離が急接近
  - 何を考えながら過ごしているか？

Product としての LLM の普及

- LLM に対する様々な使い方と日々の評価
  - 研究・開発に限らない人々が**あらゆる方法**で触る (評価する)
  - テレビ等でも当たり前のように取り上げられている

究極のマルチタスク学習

- 様々な**ユーザーの用途に対応**する必要性
  - 究極のマルチタスク学習

- 従来のマルチタスク学習研究との違い
  - 想定するタスクの数: O(1)～O(10) → 無数のユーザー入力
  - 「タスク」という概念自体の曖昧性 (何がどんな形式で聞かれるのか、どう出力すべきか、の想定が難しい)

研究としての難しさ (おもしろさ)

- 一般的な研究サイクル
  - 問題提起 → 解決法の提案 → 改善 (→ 論文)

- 追加で考えること (コアの **LLM 学習**自体への貢献として)
  - その他・既存の学習データとの共存が大事である
  - 特定の問題を解決した結果、他に悪影響がでるかもしれない

研究者としての身の振り方

- Publication の優先度の (一時的な) 低下
  - 内部インパクト (LLM 貢献度) が大きい = 秘密事項が多い
  - おもしろいコアの部分に近づけば近づくほど、対外発表をひとまず度外視する傾向になる

- どうやってうまくバランスをとるか、が (個人的な) 課題
  - こういった業界に興味を持たれている方にはここが一番重要なポイントかもしれません