＊ 質問の文言を整理している場合があります。ご了承ください。

河原 大輔 先生「日本語に強い大規模言語モデルの開発のためのコーパス構築」		
	日本語情報として戦前戦後の学術書はコーパスとして用いることは現実的に可能でしょうか。 無理だと思われる場合、最大のブロッキング要素はなんだと思われていますか。	可能と思います。 現在、国立国会図書館デジタルコレクションに収められている著作権保護期間満了の図書の利用を進めており、戦前戦後の学術書が含まれていると思います。 事前学習コーパスの質、多様性を高めることは LLM の性能に直結しますので、このような取り組みを今後さらに進めたいと考えています。
	コーパスに、書籍データ活用の可能性はありますか。 コーパスの更なる充実へ向け、出版社との連携の可能性はありますか。	現在、国立国会図書館デジタルコレクションに収められている著作権保護期間満了の図書の利用を進めています。 さらに多様な図書を利用したいと考えており、著作者の権利を尊重しながら、出版社などとの連携を検討中です。
	翻訳アプリや LLM での翻訳をしていて気になったのですが、日本語の「主語などを省略する事がある」や「色や味覚表現の多様性」などの、英語等との「言語文化の違い」に関する「翻訳」については、どのように対応していますか。	言語や文化の差による言語表現の違いは、明示的に対応しているわけではありませんが、事前学習コーパスやインストラクションデータに含まれる対訳テキストにおいて表現されていると考えられます。 翻訳の精度をさらに向上させるためには、対訳テキストを含むコーパスの規模を拡大していくことが一つの方法であると思います。
	事前学習の最終盤に特に高品質なデータや、指示形式に近い合成データを集中的に学習させるといった、いわゆる「カリキュラム学習」について何か検討されたことはありますか。 もしくは、カリキュラム学習の有望性などに関して、現時点でお考えはありますか。	カリキュラム学習は有望だと考えており、継続的に検証を行っています。 ご提示のように、高品質なコーパスを最終盤の学習に用いることや、合成データの利用を検討しています。 これらの検証結果を今後の LLM 勉強会で報告するとともに、よい結果を得た方法をモデル学習に適用し、公開する予定です。
	LLM 生成による合成データの使用は、どのような状況でしょうか。	事前学習コーパスやインストラクションデータを合成し、LLM 学習に用いることを現在検証しています。これらの検証結果は今後の LLM 勉強会で報告いたします。
宮尾 祐介 先生「大規模言語モデルのチューニングと評価」		
	マルチターンデータのチューニングの評価手法はどのようになさっていますか。 シングルターンの評価手法と同じでよいと思えないのですが、ではどのような評価指標が追加で必要とお考えか、私見でも結構ですので示唆いただけるとありがたいです。	MT-Bench などのマルチターンデータも評価に用いていますが、評価手法はシングルターンの評価と今のところ同様で単純な LLM-as-a-Judge を用いています。 対話とマルチターンチャットは正確には異なりますが、対話システムの研究における評価方法などを参照してよりよい評価方法を検討する必要があると考えています。
	評価結果としてよく出ているグラフの横軸は、学習用データセットのトークン数と記載していましたが、データセットのサイズをいろいろ変えながら、それぞれ学習して評価結果を出しているのでしょうか。それとも、学習曲線のようなものでしょうか。	横軸は事前学習のデータサイズで、事前学習の途中時点のモデルを保存しておき、それに対して llm-jp-eval などによる評価を行っています。 データサイズが異なるデータで最初から学習するわけではなく、学習途中のデータを保存しておけば評価ができるので、それほどコストがかかるわけではありません。
	チューニングと評価はセットになっているとのことですが、評価結果によってハイパーパラメータや事前学習の仕様の見直しが必要になることはあるのでしょうか。	当然ありえます。 実際、172B モデルの事前学習において評価結果に問題があることがわかり、調査の結果、ハイパーパラメータが原因であることがわかったので、新たなハイパーパラメータで学習を行うことになりました。 チューニングにおいても、評価結果を参照しながらより効果的なチューニング方法やチューニングデータを検討しています。

＊ 質問の文言を整理している場合があります。ご了承ください。

関根 聡 先生「大規模言語モデルにおける安全性の実現と方向性」		
	安全性や偽情報を守らせるために現在、テキストの条件的に判断させていると思いますが、抜本的に倫理観、心として AI に認識させる方法や可能性は感じられていますか。 テキストで条件をつけて、カウンターで打ち消して、矛盾をなくしていくことには限界が来るのではないかと感じています。 時には嘘をついたり、傷つけたり、言ったことを忘れる人間の凄さを改めて感じています。	人間の知性は本当に素晴らしいと思います。LLM はその対応を外から見ていると、一見、心を持っている様にも見えますが、大量の文章を丸覚えしてつなぎ合わせてそれらしい回答をしているだけですので、私は心は持たないと思っております。 例えば「人に迷惑をかけてはいけない」と子供に教えると、場面場面で人の気持ちを読み取り、迷惑をかけないような対応をしますが、その発言だけで、LLM が人間のように理解して人に迷惑をかけない発言をその後一切しなくなることはないと思っておりますし、現状、そのような出力しか得られていません。
	国際的に LLM の安全性を考える場合に、日本語 LLM の安全性のみだとカバーできない部分があれば教えていただきたいです。	「日本語」と「日本文化」という意味でも違うのですが、言語間の問題としては、GPT なども日本語の安全性には弱いという印象があります。 英語で言ったら新聞に取り上げられてしまうようなことも、日本語で LLM に聞くと発言してしまうこともあります。 文化的には、やはりそれぞれの国で、法律や倫理観は全く異なりますので、何が安全かということは、それぞれの国や文化として教えないといけないと思われます。
岡崎 直観 先生「マルチモーダル基盤モデル」		
	マルチモーダル AI の汎用性あるいはドメイン限定性はどの程度あるとお考えでしょうか。 たとえば、医療応用、自動運転応用、経済分析応用と考えた場合、共通の汎用モデルを微調整するようなイメージでいればよいのか、他のアプローチが考えられるのか、お知らせいただけましたら幸いです。	LLM と同様に、まずは汎用的なマルチモーダル基盤モデルを目指し、その後ドメインに特化させていく方が効率がよいと思います。必要な性能に対して、どのくらいドメインの学習データがあるのかにもよりますが。
	医療特化 LLM について、医学文献や医師国家試験からコーパスを構築していてとても興味深いなと思いました。 今後は医学文献のコーパスから学習させてマルチモーダル試験コーパスで評価するのでしょうか。	既存の医療ドメイン特化のマルチモーダルモデルも、医療に関する学術文献で継続事前学習を行い、医師国家試験などのベンチマークで性能を測定しています。
	顔文字やアスキーアートなど、テキストの視覚的表現は、マルチモーダルモデルでどのように処理されるのでしょうか。 他の視覚情報と同様に扱われますか。	面白い視点をありがとうございます。 現状は顔文字の形状などが考慮されている訳ではなく、顔文字に対して埋め込み表現が割り当てられている状況だと思います。これは漢字に対する取り扱いと同様で、へんやつくりを考慮した取り扱いになっていないのと同じです。
	ガートナーのハイプ・サイクルの紹介で、ご言及のあったところではありませんが、まだ黎明期となっている部分に「エッジ生成 AI」の項目があったかと存じます。ハードウェアのスペック向上に伴い「エッジ AI」として想定される範囲も変わってくるかと思いますが、今後、「エッジ生成 AI ならでは」の研究はどのような方向で進むと考えられますか。	モデルの圧縮（既存のモデルのメモリ容量の削減）、推論の高速化、小さいモデルをフルスクラッチで学習する試みなどが考えられると思います。 この辺が参考になると思います。 https://edge-llms-challenge.github.io/edge-llm-challenge.github.io/index