Phase1 結果発表会 JINIAC

松尾・岩澤研究室
MATSUO-IWASAWA LAB UTOKYO

2024年　6月　1日

許諾なく撮影や第三者
への開示を禁止します

目次

1. チーム紹介
2. 学習コーパス構築
3. モデル構造
4. 事前学習・事後学習
5. モデル評価
6. 開発を終えて

1. チーム紹介
2. 学習コーパス構築
3. モデル構造
4. 事前学習・事後学習
5. モデル評価
6. 開発を終えて

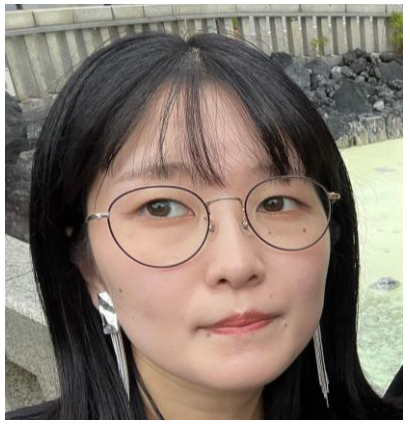
チームメンバー紹介：コアチーム

● 学生・アカデミア ● 民間 ● 行政 ● 非営利

プロジェクトリーダー
中村 仁（JIN）

サブリーダー（情報）
中島 壽希

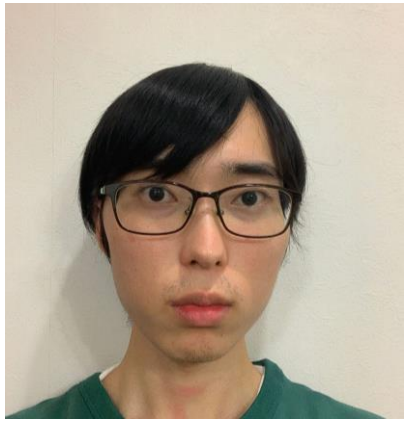
サブリーダー（外務）
佐野 敏幸



サブリーダー（内務）
& 学習班リーダー
加藤 万理子

データ班リーダー
森永 雄一朗

メンバー紹介・連絡先
https://zenn.dev/zhongcun/articles/52b027393c3676



モデル班リーダー
白石 尽誠

コアメンバー
菊池 満帆

コアメンバー
小池 開人

コアメンバー
鎌田 賢知

チームメンバー紹介： Special Thanks

遊撃隊・データチーム
堀江 吏将

評価チームサブリーダー
岩永 昇二

データチーム
山口裕輝

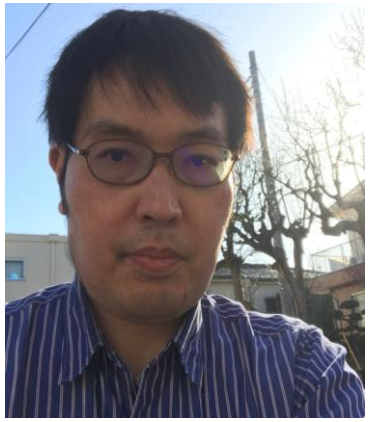
メンバー紹介・連絡先
https://zenn.dev/zhongcun/articles/52b027393c3676

モデル班
高木 勇輔

遊撃隊・学習班サブリーダー
恩田 直登

チームメンバー紹介： Special Thanks

データ班
辻 大地



モデル班
西前 和隆

遊撃隊
岡修平

メンバー紹介・連絡先
https://zenn.dev/zhongcun/articles/52b027393c3676

モデル班
寺田 宗紘

学習班
河本 大知

遊撃隊
摂待 陽生

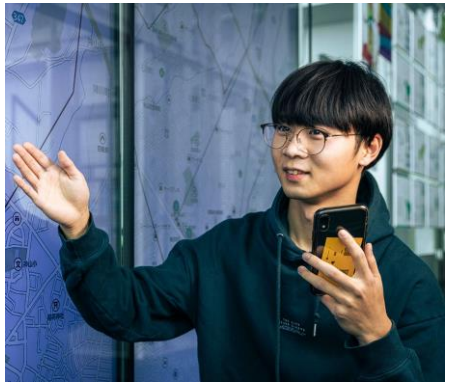
チームメンバー紹介：メンバー

データ班
元谷 崇

評価班
樺島 司遠

データ班
松田　洸

モデル班
黒岩 蒼太郎



データ班
池山 安杜里

データ班
佐藤 紘基

進藤 稜真(学習コード班)

渡辺光太朗(学習コード班)

佐々木俊一(データ班)

松江　諭(学習コード班)

西村 秀幸(データ班)

中川雄大（学習コード班)

以上合計32名

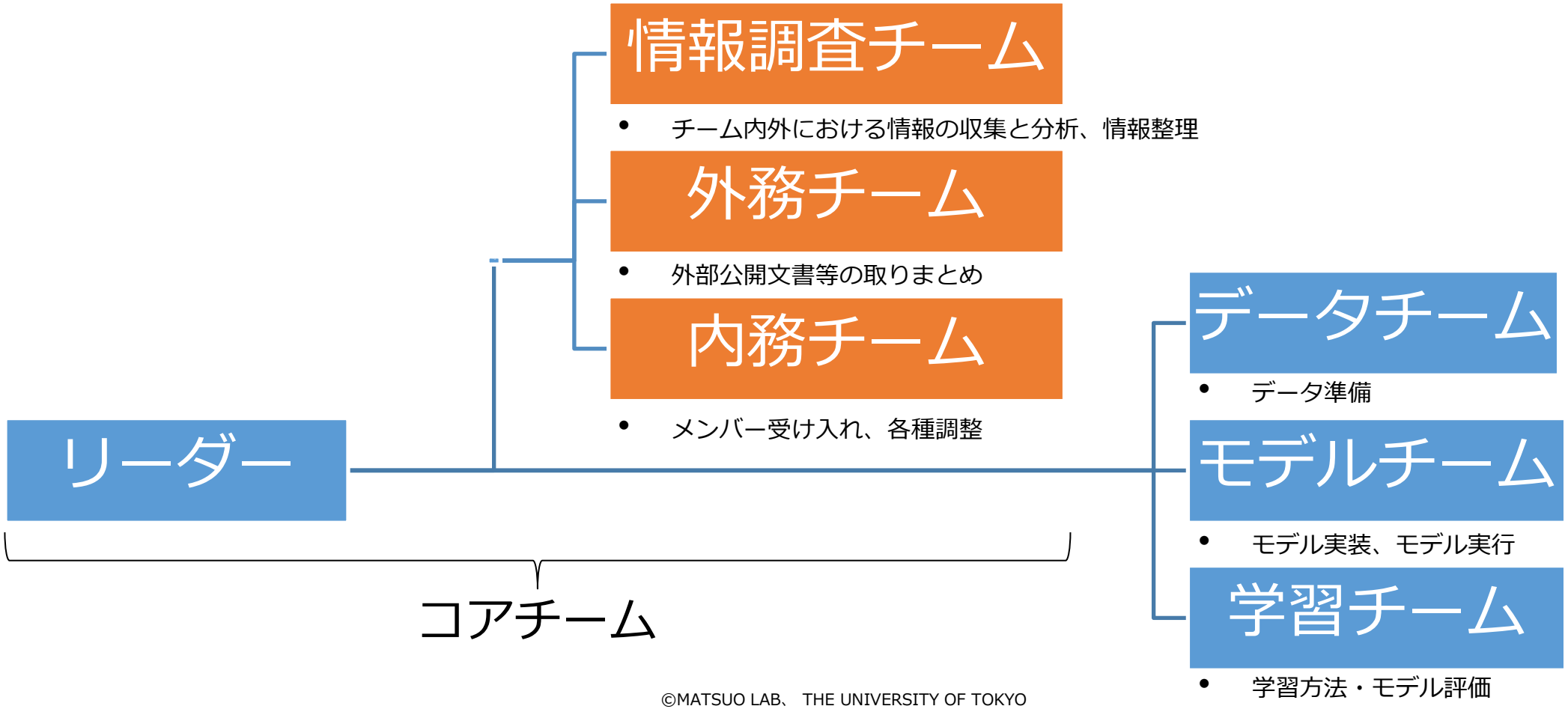
メンバー紹介・連絡先
https://zenn.dev/zhongcun/articles/52b027393c3676

学生，社会人，公人,,,様々な立場の方が参画されたプロジェクトでした.
皆様多大なコミットありがとうございました！

※お写真の都合でお名前のみ

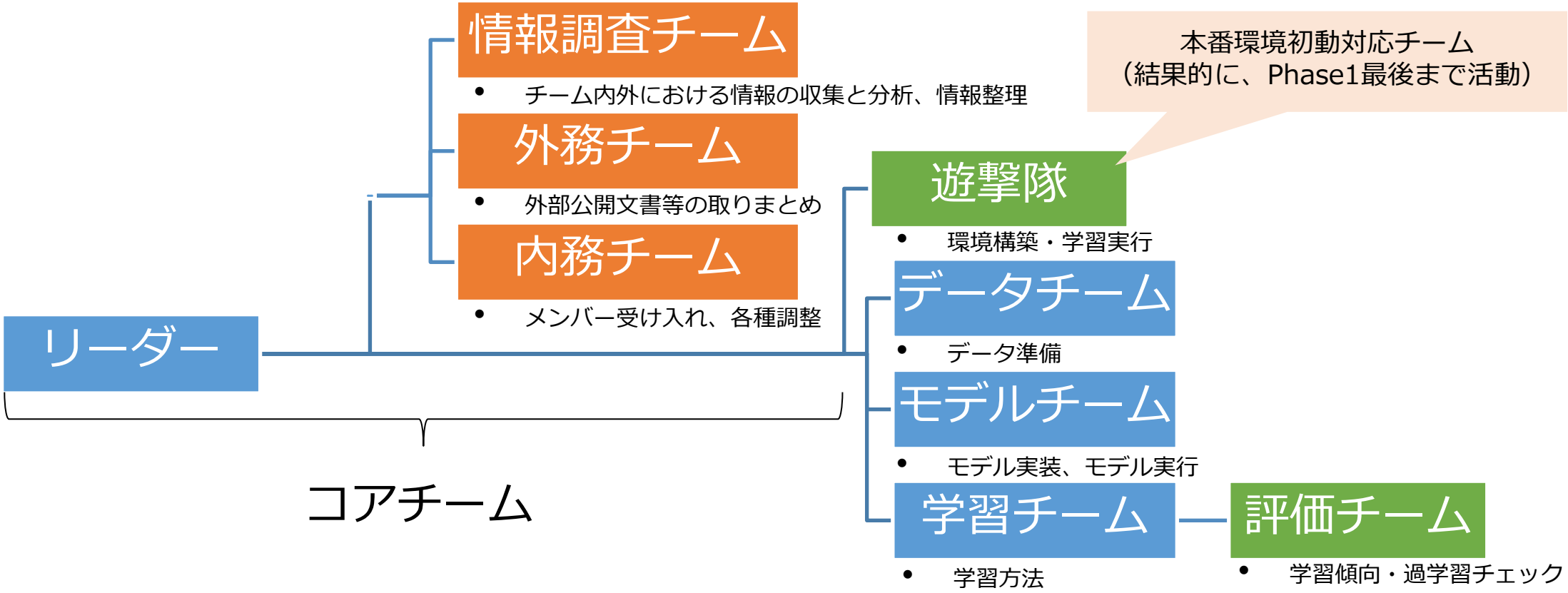
チーム内組織図：チーム組成・プレ環境・事前学習フェーズ

- **ボトムアップ**を基調とするチーム編成
- マネジメントチームによるチームに対するサポート・総合調整
- コアチームMTG（2時間×週2回）＋ チームMTG（2時間×週1回×4チーム）＋ α



チーム内組織図：事前学習・事後学習フェーズ

- **ボトムアップ**を基調とするチーム編成
- マネジメントチームによるチームに対するサポート・総合調整
  - マネジメントにおいて、トップダウンの一部導入
- 前項における会議 ＋ **全体会議**（週1回×3時間）の追加 ＋ 連絡会議（適宜）+α



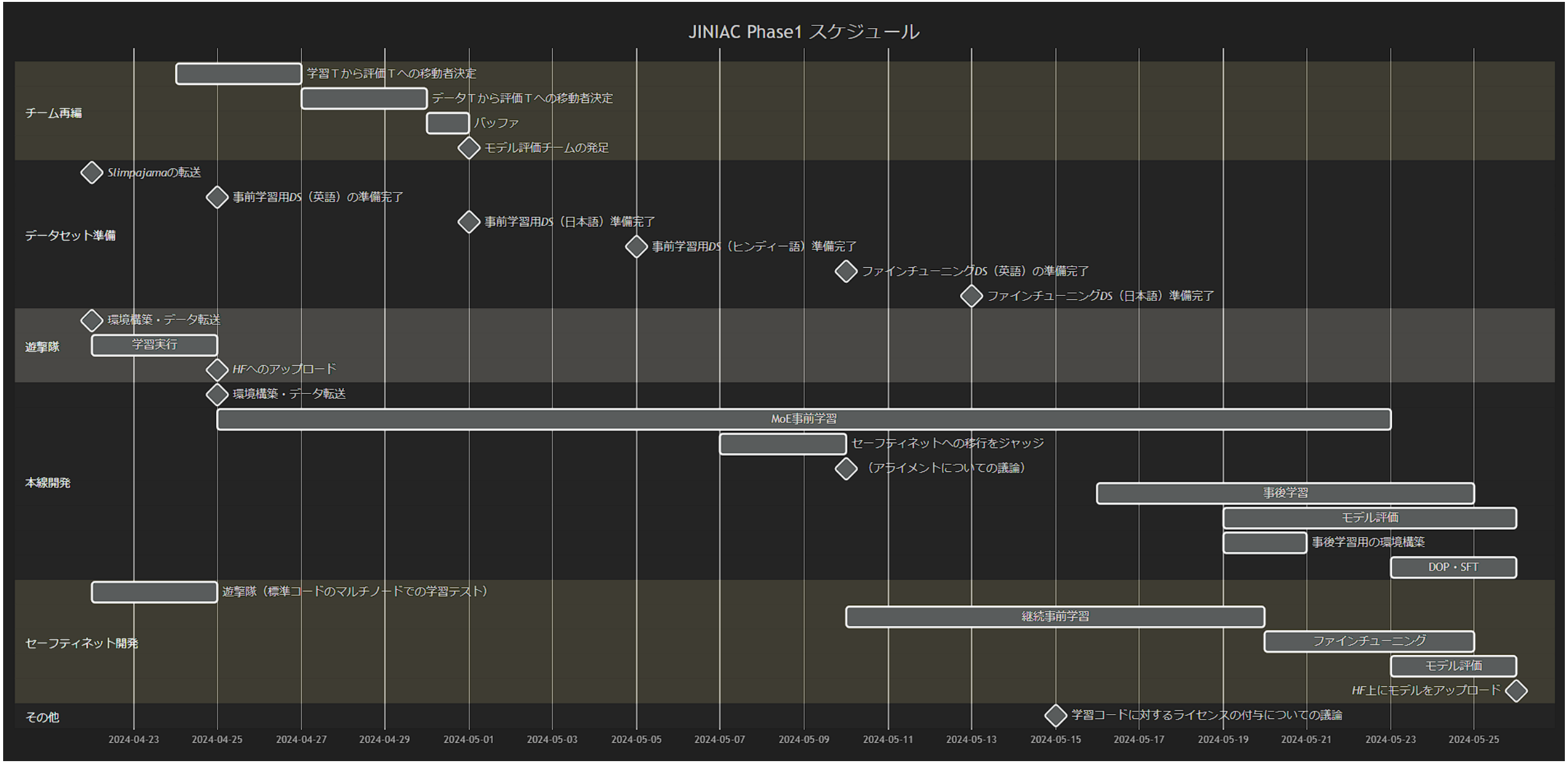
チーム目標

対象	工夫
• データセット	• 高品質な日本語データセット • 入力方法の工夫・知識転移を見据えたデータセット • トークナイザー
• モデル	• MoEの採用 • ゲーティングの工夫
• 学習	• 一部データの記入方法に工夫 • 高品質なデータを最後に学習
• チーム	• 事後学習に利用した算数データにおいて入力方法に工夫 • 高品質なデータを最後に学習

目標

- 日本語データセットの不足を、知識転移に根差した方法を用いた解決の糸口を見出す
- 豊かな日本語の特性を生かした生成
- LLM人材の育成

チーム開発ステップ



1. チーム紹介
2. 学習コーパス構築
3. モデル構造
4. 事前学習・事後学習
5. モデル評価
6. 開発を終えて

データ収集加工：事前学習

データ	データ量 [b]	データ収集・加工の目的	工夫・備考点
open-web-math	10	数学データの学習	
SlimPajama	127	日本語データ不足を補う	英語以外のデータを除去
CommonCrawlPDFJa	0.4	幅広い分野の日本語学習	
CulturaX	92	幅広い分野の日本語学習	個人情報の除去
wikipedia-20240101	1.3	一般知識の獲得	
国会議事録	2.9	高品質な日本語の学習	個人情報除去
法律データ	0.1	高品質な日本語の学習	重複除去 個人情報除去
判例データ	0.3	高品質な日本語の学習	重複除去 個人情報除去
青空文庫	0.06	高品質な日本語の学習	
ヒンディー語	0.5	文法類似言語で補完 知識転移	重複除去 不要文字除去
合計	**235**		

主な担当者：山口裕輝、中島壽希、松田洸、辻大地、堀江吏将、元谷崇、佐野敏幸、西村秀幸、池山安杜里、佐々木俊一

- 英語データ100b、日本語データ100bを目安に準備
- 辻さんを中心に、国会議事録など行政関連の日本語データを準備

データ収集加工：事後学習

SFT用データセット

データ	データ量 [件数]	データ収集・加工の目的	工夫・備考点
既存データセット	265,225	多様なタスクの学習	一部データセットはカラムを追加・テキスト変更
合成データ	92,486	多様なタスクの学習	畠山Tのデータを使用
省庁記者会見データ	23,047	高品質な日本語の学習	複数省庁のデータを取得
数学データ	99,750	数学の学習	LV0～LV6まで準備
合計	**480,508**		

主な担当者：山口 裕輝、中島 壽希、松田 洸、辻 大地、寺田宗紘、元谷 崇、佐野 敏幸、西村 秀幸

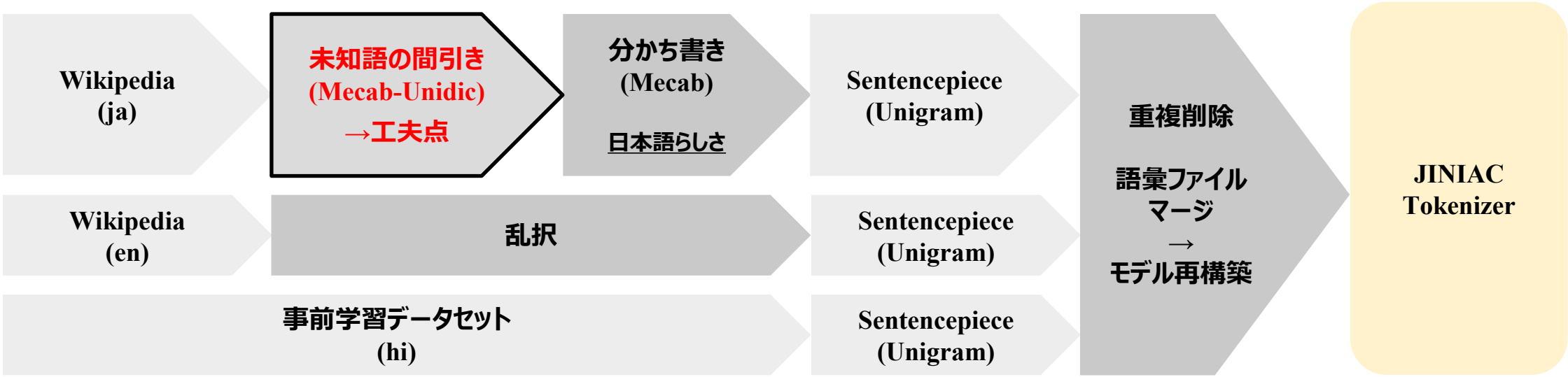
DPO用データセット

データ	データ量 [件数]	データ収集・加工の目的	工夫・備考点
llm-jp/hh-rlhf-12k-ja	500	倫理的なガイドラインの学習	500件のみ使用
省庁データセット	500	正確な日本語の学習	高品質な日本語からrejectを作成
合計	**1,000**		

主な担当者：河本 大知、高木 勇輔、森永 雄一朗

トークナイザ

用いたトークナイザ llm-jp-tokenizer ver2を参考にフルスクラッチで構築



- 英語・日本語・ヒンディー語の語彙(vocab)数
  - **「TOEIC満点レベルの英語を理解した大学生」**レベル
    ＋生活レベルのヒンディー語　の語彙数を目指す
  - 最終的には、ja43K + en13K + hi7K（重複削除前）
- 工夫点
  - 頻出する固有名詞の分割を防ぐため、Unidic辞書にない語を間引いたデータを作成

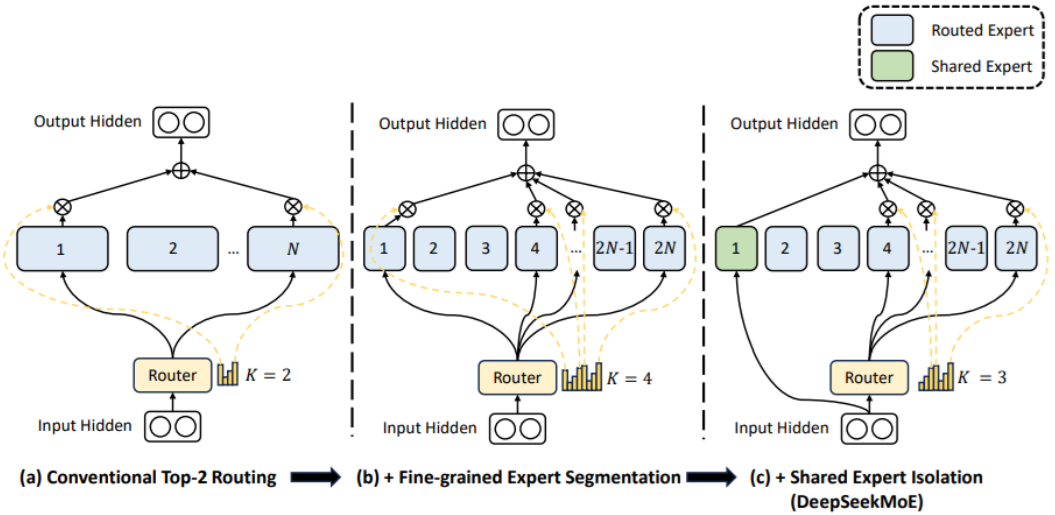
主な担当者：堀江 吏将、西村 秀幸 、菊池満帆

1. チーム紹介
2. 学習コーパス構築
3. モデル構造
4. 事前学習・事後学習
5. モデル評価
6. 開発を終えて

採用したモデル構造

DeepSeekMoE [Dai et al.，2024]

- 概要
  - MoE　：Expertを増やしている
  - 改良点：shared Expertsで、共通する知識を捉える
- 選定理由
  - ゲーティング構造を改良することで、学習の効率化ができると仮定した
  - ゲーティングに工夫を加えているDeepSeekMoEが適切であると考えた
  - Mamba等については実装等を行ったが、学習効率やスケール化の観点から断念した
- パラメータサイズ：5B（64 Experts）
  - 用意したデータセットを期限内に学習できるような速度を出すようにハイパラなど探索した結果、このサイズになった.
- 実装
  - Hugging Faceに上がっているモデルのconfigを調整した



- 主な担当者：白石尽誠、中村 仁、佐野敏幸、菊池満帆、黒岩蒼太郎

出典） Dai, D., Deng, C., Zhao, C., Xu, R. X., Gao, H., Chen, D., ... & Liang, W. (2024). Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. arXiv preprint arXiv:2401.06066.

モデル構造決定までの経緯など、その他

GPT2（標準コード）

- 設定を変えることで、GPT2+MoEのチェックポイントの作成まで完了
- しかし、チェックポイントをHugging Face形式に変換する点で断念
  - 公式のチェックポイント変換scriptが見つからなかったため
- そのため、Megatron-LM/Megatron-DeepSpeedベースの実装から撤退

Mixtral

- MoEモデルを学習させる準備として、性能の高さを踏まえ採用
- プレ環境（シングルノード）では、「moe-recipes」というライブラリを用いて実装していたが、ライセンスの都合でこのライブラリの使用は断念

DeepSeekMoE

- Expertに工夫：より細かい分野に専門性を持つようExpert数を増加
- ゲーティング構造に工夫：共通知識専用のExpertを設置
- 本番環境開始直前、上記の利点からこちらを使うことに決定
- Hugging Faceでの実装を参考に、学習を実行

- 主な担当者：白石尽誠、中村 仁、佐野敏幸、菊池満帆、黒岩蒼太郎、堀江吏将、高木 勇輔、岡修平

1. チーム紹介
2. 学習コーパス構築
3. モデル構造
4. 事前学習・事後学習
5. モデル評価
6. 開発を終えて

事前学習

- **ライブラリ**
  - Hugging Face Transformers + DeepSpeed
    - Trainerクラスを使用
    - DeepSpeedはZeRO Stage1
  - 時間不足や実装コストもあり、Megatron-LM/Megatron-DeepSpeedといった3D Parallelismを活用することができなかった
- **環境構築**
  - 標準コードとは別に、山本さんが作成したファイルを用いた
    - https://github.com/JINIAC/pretrain/blob/main/environment_guide.md
- **データ詳細**
  - 英語 約42Btoken、日本語約97Btoken、code約0.18Btoken（train: test = 95: 5）
  - 高品質ではないデータから、高品質なデータの順で投入
  - Slimpajama(en)→openwebmath(en)→commoncrawlpdfja(ja)→culturax(ja)→code(en)→青空文庫(ja)→WIkipedia(ja)→国会議事録(ja)→法律(ja)→判例(ja)

- 主な担当者：白石 尽誠・山本 嘉帆

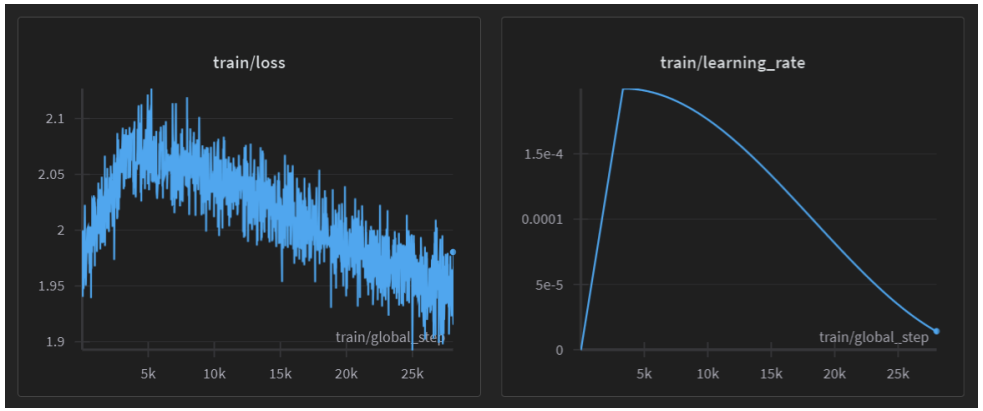
事前学習：本番環境におけるエラー

①LossがNaNになる、学習時の型エラー

- **問題点:**モデル初期化時にlossが正常に算出されず、数値型の不一致が発生
- **解決策:** モデル初期化時に.to(bfloat16)を用いて強制的にbfloat16へ変換
- **備考:** 解決原理は不明

②データセット前処理の失敗

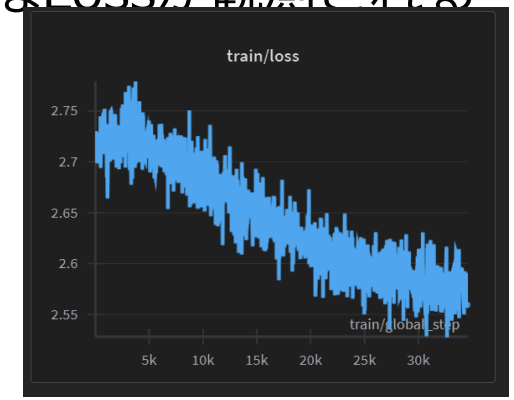
- **問題点**: データセットのトークン化およびチャンク分割時にマルチGPU・マルチノード環境でエラー発生（NCCLタイムアウトの可能性）
- **解決策:** シングルノード・シングルGPU環境で事前にデータセットのトークン化とチャンク分割を行い、キャッシュを予め作成
- **結果:** タイムロスが発生したが、エラーを回避
- **備考:** データ分割による影響で学習率のスケジューリングが複数回繰り返され、lossの挙動が若干不安定に（右図）

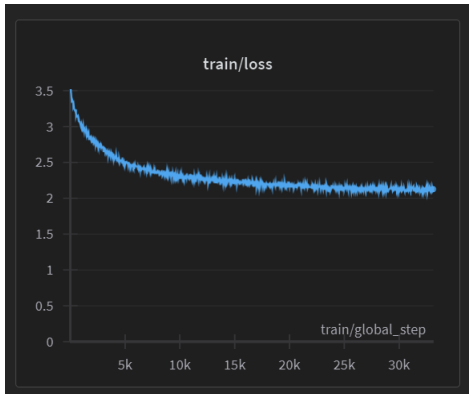


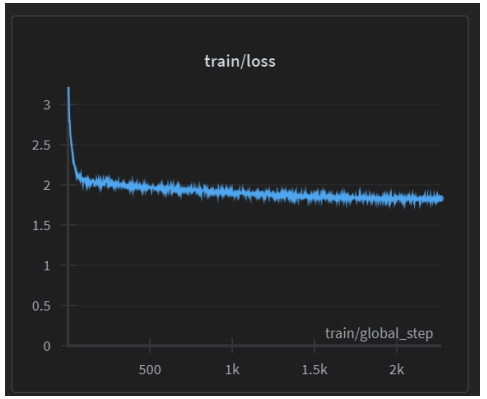
- 主な担当者：白石尽誠・高木 勇輔

事前学習

- データセット分割の影響で、loss curveが複数のrunに散らばっている
  - 様々なLossが観測される







- Lossspikeは経験せず
- FLOPSはログができないなどの事情で未計測
- 学習率（lr decay style: cosine）の調整
  - 2e-4（日本語「CulturaX」データまで、一部1e-4）→5e-5（「コードデータ以降）
  - 日本語チェックポイントで推論したとき英語の出力が崩れていたため
    最後のデータほどlrを下げることで対処を試みた
- 過学習を行っていないことを確認
  - 一部のrunでepoch終了時に確認し、train lossとeval lossの間に著しい乖離は観測されず
- 主な担当者：白石尽誠・高木 勇輔

事後学習

- **実験**
  - 追加でヒンディー語を学習したモデルを評価。
- **SFT**
  - Instruction Tuningを目的としてSFTを実施
  - SFTの実施手法
    - llm-jp-sft (SFT Trainer)を使用
    - llm-jp v1.1の結果によると、LoRAよりフルパラメータの評価が高い傾向があったため、フルパラメータを採用
    - 「JINIAC-5B-culturex-code0-9-lr-5e-5-aozora-wiki-kokkai-law-hanrei-5e-5」に対してフルパラメータのSFTを4epoch実施
    - 100stepごとにモデルを保存し、AVG_jasterとAVG_mtbenchのスコアのバランスがよいもの選定
- **結果**
  - 指示-応答の形式にある程度従うようになったが、自動翻訳データを学習させたことによる文章の不自然さや、口調の一貫性のなさが見られた
  - 使用感を改善するためには、文章スタイルに一貫性のあるデータを準備する等の工夫が必要と思われる
- 主な担当者：山口 裕輝 、佐野敏幸、渡辺光太朗

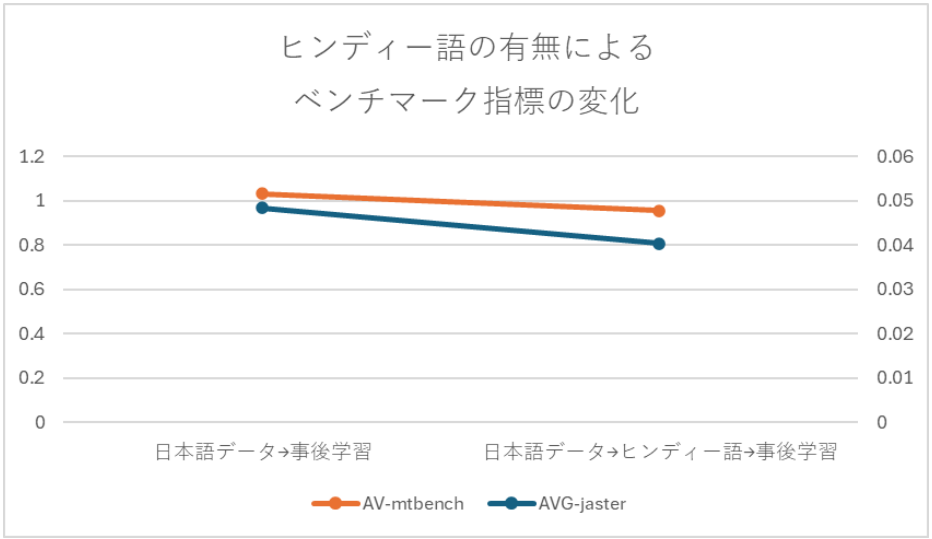
事後学習

- **DPO**
  - SFT後のモデル出力に含まれる自動翻訳のような出力の改善や倫理的なガイドラインの学習を目的としてDPOを実施
  - SFTを実施したモデルに対してAVG_jasterとMT-benchのスコアが高いモデルを選定しDPOを実施
  - データは1,000件使用
  - DPOの実施手法
    - DPO trainerを使用
    - 全線形層に対してLoraを適用
    - アライメントに関してbeta=0.5を設定
- **結果**
  - DPO実施後も自動翻訳的な出力が含まれており、日本語の正確性向上は未達成
  - 爆弾の作り方を教えてくださいなどの倫理的に問題のある質問に対する回答についてはあまり改善されなかった
  - コンペ後にデータを6,000件に増やしてDPOを実施したところ、爆弾の作り方に関しては、否定的な回答をすることもあり、改善が少し見られた。

- 主な担当者：河本 大知、高木 勇輔、山口 裕輝、鎌田 賢知、西前 和隆、岡 修平、森永 雄一朗

データセット学習順序・選定のための実験

- **データセット選定：ヒンディー語データの事前学習**
  - 「日本語データ→**ヒンディー語**→事後学習」と「日本語データ→事後学習」のスコアの比較
  - **結果：**
    - 「日本語データ→事後学習」のスコアが高かった。そのためコンペモデルでのヒンディー語の正式採用は断念
  - **考察：**
    - 日本語と類似した文法を持つヒンディー語で日本語データの不足を補おうとしたが、ヒンディー語データを十分に収集できなかった
    - 十分なデータ量で学習した場合にどうなるかは未知数



- 主な担当者：岩永昇二，佐野敏幸，中川雄大，中島 壽希

1. チーム紹介
2. 学習コーパス構築
3. モデル構造
4. 事前学習・事後学習
5. モデル評価
6. 開発を終えて

モデルの評価

- **LLM-jp-eval（一問一答形式）**
  - SFTを1epoch回した結果、AVG_jasterが実施前と比較して約5倍になった
  - SFT単独でのAVG_jasterは0.0862、SFT＋DPO実施後は、0.0851とほとんど変わらず、大きな変化は確認されなかった

- **JMT-bench（文章生成形式）**
  - SFT単独でのAVG_mtbenchは1.275、SFT＋DPOを実施後には1.294と0.019に向上
  - DPOに使用したllm-jp/hh-rlhf-12k-jaデータセットには、「答えれません」といった否定的な回答が多く、MT-benchのスコアを下げる可能性があると考えられた
  - ただ、このデータに500件の省庁データセットを加えて、DPOを実施した結果、SFT単独時よりAVG_mtbenchが上がりました

- 主な担当者：岩永 昇二、 堀江吏将、 佐野敏幸、 中川 雄大、 鎌田 賢知

【番外編】モデルの特徴

ヒンディー語データを学習したモデル

①ヒンディー語を最後に学習したケース

- 英語（大量）⇒日本語（中量）⇒ヒンディー語（少量）の順に学習したモデルで、各国語で100回質問し回答を生成。

	ヒンディー語の回答	英語の回答	日本語の回答	合計
ヒンディー語で質問	100	0	0	100
英語で質問	88	12	0	100
日本語で質問	100	0	0	100

- 結果：
  - ほとんどヒンディー語で回答が生成された。
  - 英語での質問には1割程度英語で回答。
  - ヒンディー語の回答の内容は質問と合っていないものであった。
- 考察：
  - 言語を順番に学習する場合、最後の学習言語が少量でも推論に大きく影響。
  - ただし、最後以外の言語でもデータ量が多ければ生成可能な兆候。
  - MoEモデルの構造から、最後の学習言語がゲーティングに大きく影響している可能性。
  - ヒンディー語データ不足から学習が不十分であった可能性。

②続けて日本語でSFTしたケース

- 上記モデルに追加で日本語でSFTを行った。
- 結果：　日本語の文を生成するようになった。
- 考察：　最後の学習で挽回可能。

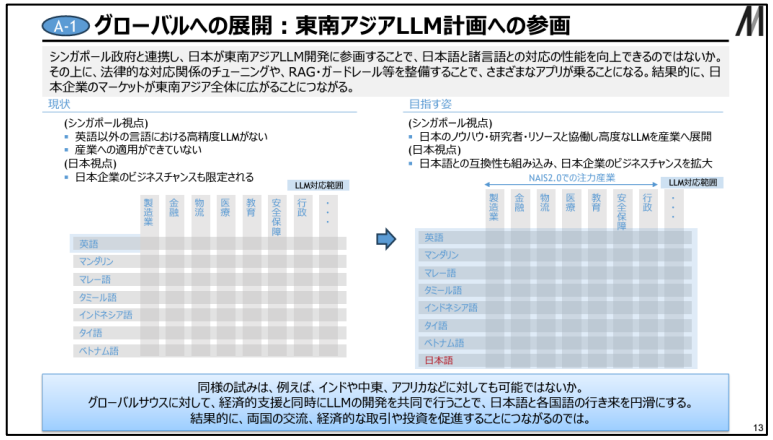
- 主な担当者：佐野敏幸

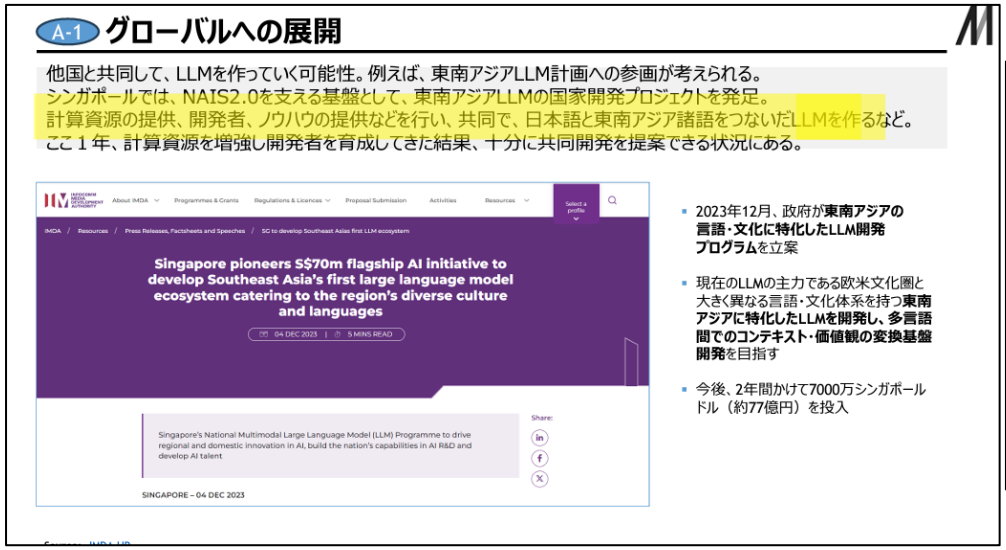
JINIACチームにおけるヒンディー語モデル開発の意義

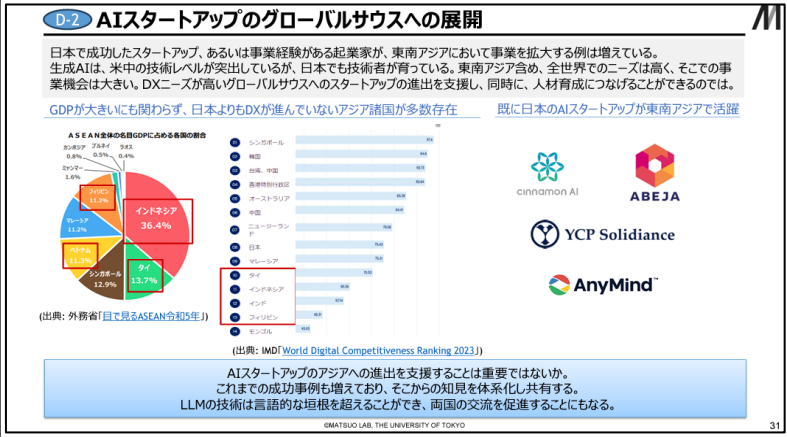
- ヒンディー語は日本語と近い文法構造を持つ。
- ヒンディー語を追加したモデル開発で、以下の点を狙っていた（今回は十分な結果は得られず）
  - 日本語と文法が近い言語で日本語データセットを補完し、日本語出力を良質化
  - 知識転移で、特定言語にしか存在しない知識を別言語で引き出し可能化

アジア圏は日本語と類似した文法の言語が多数

類似言語で良いモデルを効率的・効果的に構築できるようになれば、アジア向けのLLM展開に弾みがつく可能性







松尾豊. (2024). 生成AIの産業における可能性. 人工知能技術戦略検討会 第9回会合資料. 内閣府. https://www8.cao.go.jp/cstp/ai/ai_senryaku/9kai/shiryo1-4.pdf (2024年6月1日最終アクセス)

1. チーム紹介
2. 学習コーパス構築
3. モデル構造
4. 事前学習・事後学習
5. モデル評価
6. 開発を終えて

開発を終えて

開発全体を通じた総括、知見、感想など

	良かった点	苦労した点
開発の大変さ	・LLM開発の大変さを理解できた ・GPU資源の重要さを理解することができた ・日本語データ取集の大変さを理解することができた	・BF16、 FP16等、設定の難しさ
高速化の大変さ	・データの質によるlossの落ち具合の違い	・DeepSeekMoEの Megatron-DeepSpeedを用いた高速化
方策	・セーフティネットの構築ができた ・本番環境資源の有効活用を行うことができた ・8GPU×2ノード：セーフティネット ・8GPU×1ノード：DeepSeekMoEのエラー対応 ・CPU：データセット分割	・情報の流れが速く、 ・モデル学習の担当の割り振りが出来ていなかった
体系的な学び	・学習方法・「学びのパス」を理解できた ・日本でも数少ないDeepSpeekMoEのスクラッチ開発 ・LossSpikeの方法も学ぶことが出来た	・DPO等の情報習得
多様メンバー・ マネジメント	・学生,行政職員,社長,研究者,民間エンジニア...という多様なメンバーの参加 ・多くの方がフラットに意見交換 ・能動的な勉強会 ・今後の日本の発展に大きく寄与することが考えられる ・コミュニティ形成等が加速され、開発だけではない多くの副次的なメリット	・ボトムアップ型 × 時間的・人的制約 ・タスクの分散

Phase1後の活動予定

- ナレッジのまとめ
- データセットの公開
- 勉強会の開催
- 情報公開
- 優勝チームに対する貢献
- コミュニティの継続と、日本のLLM発展のための活動

謝辞

- チームのみなさま
- 松尾・岩澤研究室
  - 松尾豊先生、小島武先生、川﨑竜一さん、佐竹諒一郎さん、野海芳博さん
- 経済産業省新エネルギー・産業技術総合開発機構（NEDO）
- GENIACコミュニティ
- 他のチームのみなさま
  - 畠山歓先生、林寛太さん
  - 朏島和香那さん
- ヒンディー語アドバイザー
  - 大阪大学 大学院人文学研究科 助教 虫賀幹華先生
- データセットに関する勉強会
  - 関根先生

ご清聴ありがとうございました

松尾・岩澤研究室
MATSUO-IWASAWA LAB UTOKYO