



中村チーム（仮称） GENIAC OSS LLM メンバー向けプレゼン

中村 仁

公立はこだて未来大学 システム情報科学部 B4

自己紹介・コアメンバー紹介

本チームのリーダーとコアメンバーについてご紹介します



- 氏名：中村 仁 / Jin NAKAMURA
- 専攻：複雑系科学
- 研究：脳型AIを用いた行動計画のモデル
- 学会：日本神経回路学会（JNNS）
- 松尾・岩澤研究室 / LLM関連活動：
 - 松尾・岩澤研究室 LLM Summer 2023 11位
 - 松尾・岩澤研究室 LLM-DXインターン（昨年12月～）
 - 地方公共団体のオープンデータ取組支援（昨年12月～）
 - 松尾・岩澤研究室「知能を創る」プロジェクト（3月～）
- #05_自己紹介 における自己紹介：



研究・プロジェクト経験

● 研究

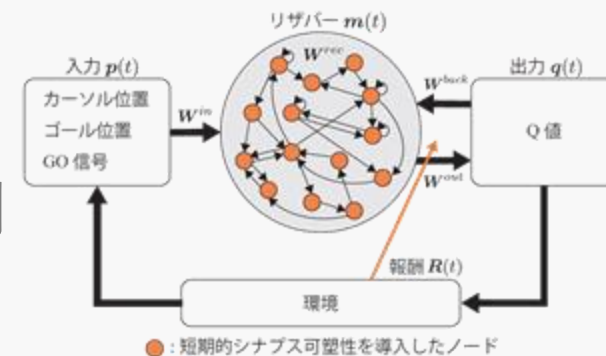
- 動的シナプスを用いた報酬修飾型リザーバー計算に基づく行動計画の数値モデルと、脳内における情報表現の解明

● 学会・実績

- JNNS2021, 「多細胞バイオ」第2回領域会議, NEURO2022, BFBC2023. (B2~B4)
 - IEEE WCCS 2024での発表、ジャーナルへの投稿を予定.
- 大学賞, 国際学会最優秀賞. (B3)

● プロジェクト

- 「脳をつくるプロジェクト」プロジェクトリーダー
「世界モデル (Dreamer) を用いた自動運転の実現」



小池 開人



- アポロ株式会社 所属
- オントロジー, 機械学習
- コンピュータと日本の言語統一を目指したいです

中島 壽希(nakashi104)



- ソフトウェアエンジニア&社会人学生 (NLP研究)
- LLMの推論能力や映像・ダイナミクスと言語の接地に興味があります
- 楽しみましょう！

菊池 満帆



- 株式会社リクルート
- LLMのアルゴリズムの最適化とAI × 労働生産性に興味あります！
- 和気あいあいと楽しくやりましょ！

鎌田 賢知



- 早稲田大学院 → 新社会人
- CV, 音響, DS, ハッカソン
- 貴重な機会なので一緒に高め合えたら最高です！

LLM開発について

本チームの注力領域に関してご説明します

真に有用な日本語LLM開発の手がかりを見つける

- 評価向上を狙いつつも囚われすぎず、
「日本語能力」向上に寄与しそうなアイデアを多く試したい
 - 面白いアイデアが見つかれば、
優先順位に配慮しつつも、班を作り、ゲリラ的に試していきたい
 - その結果として、8チーム中1位を目指す
 - 個人ではできない、ユニークな開発を行いたい
 - 最終的には、国富に寄与するLLMを作りたい
- ※詳細は、コアメンバー・メンバー相談しながら決定

本プロジェクトの主な注力領域

メンバー・コミュニティメンバー

◎データセット (DS)

- 良質な日本語DS/知識転移を想定したDS
 - 既存のDSは△
- 前処理（トークナイザー等）の適切な選定

メンバー・コミュニティメンバー

○モデル構造

- MoE
- 脱transformer：Mambaなど

メンバー・コミュニティメンバー

学習手法

- 「繰り返し学習」・
「人間的なカリキュラム学習」など
- モデル構造を鑑みながら選定

メンバー中心

○環境（GPU・高速化）

- 分散学習
- モデル構造やモデルサイズに応じた適切なライブラリ

「日本語のLLMを作る」ということについて

- **問題**：既存のLLMsは、日本語のニュアンスがおかしい



- **仮説**：良質な日本語のDSをあまり用いていないため
- **使用したいデータ**：Common crawl厳選+良質な自前データDS
 - 政府・自治体における日本語資料：白書、国会QA、自治体資料、教科書
 - 日本語会話データ：SNSデータ
 - 日本語論文：大学の学術リポジトリ
 - その他：古典

※全て、権利関係を確認中。独自制作データ制作し、公開する予定

※畠山先生のデータセット作成案に賛同しており、ぜひ協力させていただきたい

「日本語のLLMを作る」ということについて

- **問題**：良質な日本語DSが少なすぎる



- **仮説**：知識転移を前提とした言語選択し、DSを選択
- **使用したいデータ**：日本語に近い文法の言語DS
 - ヒンディ：インドに精通した官公庁関連の方々と連携して調査中（採用未定）
 - その他、韓国語、トルコ語等も検討中
- **想定される問題**：文化の差異
 - 単語の概念が異なる場合がある
 - 例：食文化により、「ゴキブリ」が「食べもの」「害虫」と異なるなど

データセットの作成でネックとなりそうな点

- 著作権
- アノテーション
 - これまでの日本語データセット(日本語訳dollyやmc4)の質が非常に悪いため、改良が必要
 - LLM勉強会でデータセットの担当されていた方（メンバー予定者）と、方法をいくつかの案を練っている最中

これに関連したネックとなりそうな点

- 日本語以外のデータセットにおけるトークナイザーの選定

事前学習

- 試行回数を増やしたいので、
モデルサイズにより適切な分散学習の設定を行いたい
- 事前学習ライブラリについての選定
 - DeepSpeedを改良する方法を考えたほうが良さそう
- その他も色々…（割愛）

チームング

このようなチーム運営をしていきます

チーム運営方針

メンバー・コミュニティメンバー

チーム目標・方針

- コアメンバーと相談
- 他のチームに対する情報公開・コミュニケーション
- 開発メンバー・
コミュニティメンバーからの意見も多く反映

メンバー中心

スケジュール

- 状況に応じて臨機応変に対応
- 様々なメンバーの事情にも配慮

メンバー・コミュニティメンバー

雰囲気

- 雑談できる感じのイメージ
- LLM Summer 2023における優秀生
- 公務員・大学生



メンバー中心

調整役

- 14h以内に返信できると嬉しい
(僕：24時間連絡OK)

どのような技能等が前提？

コミュニケーション

必要な役職の明確化

必要な技能の明確化

どのような技能等が前提？

カード①

報連相・稼働時間

- 稼働時間は時期により変動があってもOK
- 昼だけ・夜だけOK
- 会議は、週に2回ほど
 - 運営系
 - 開発系
- 会議は録画し、可能な範囲で公開

カード②

裁量権：あり

- サブリーダー・書記など
- 一定のレベルで自由に決めて頂けるようにする
 - 責任は発生しない
- 重要な内容は、運営系の会議で決定

カード③

技能：これからでOK

- 報連相・コミュニケーション・やる気！
- 開発経験は問わず
可能なら
- データセット作り経験
- モデル構造への理解
- 各種アイディア

想定しているプロジェクト管理ツール

● Notion

- メンバー情報・進捗管理・報告
- 情報が失われない様、2つのデータベースを工夫して運用

● Google Calendar (Notion Calendar)

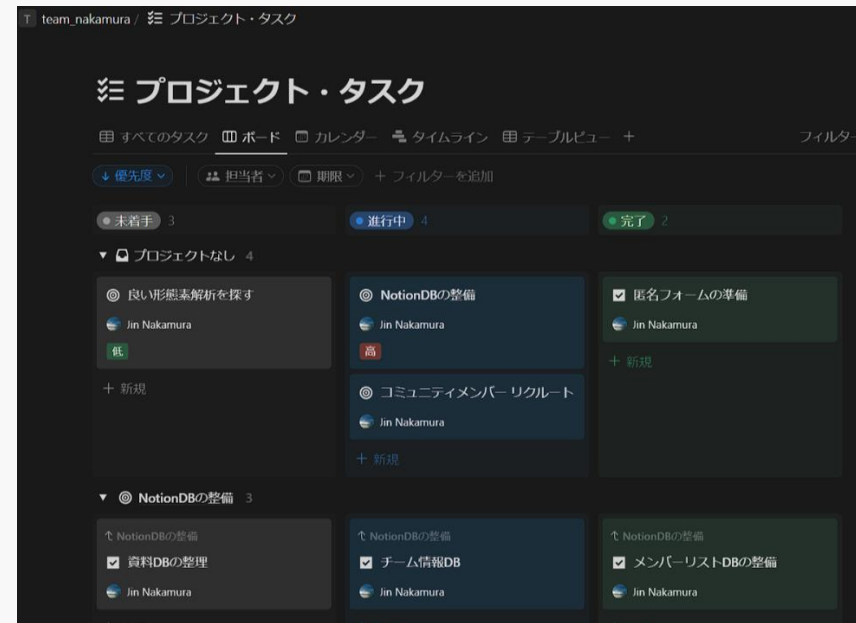
- 日程調整

● Slack

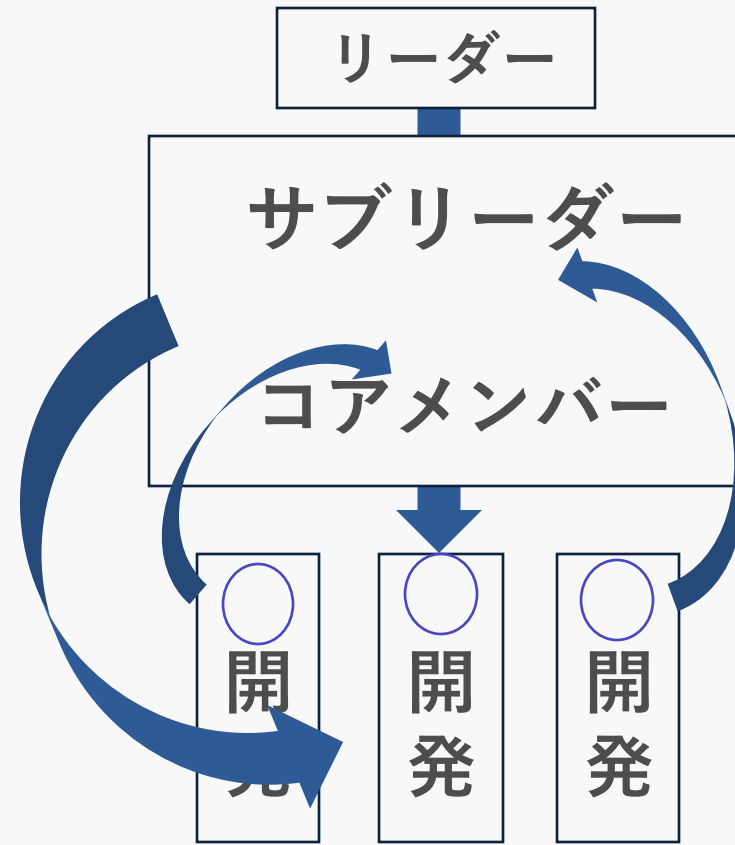
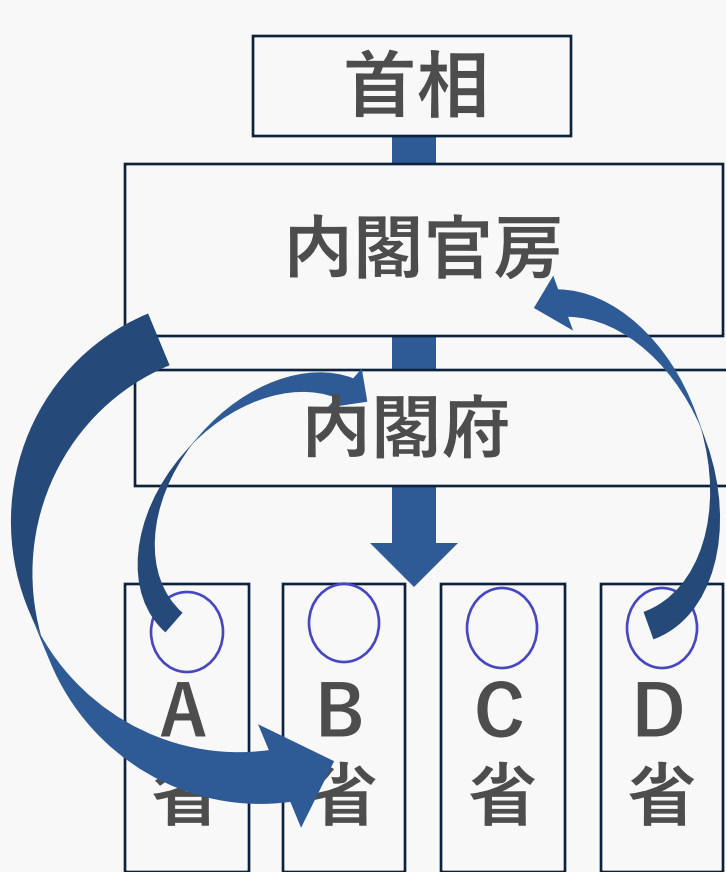
- 報告連絡相談・コミュニケーション
- カジュアルなお話は、DMグループで行う予定

● Miro

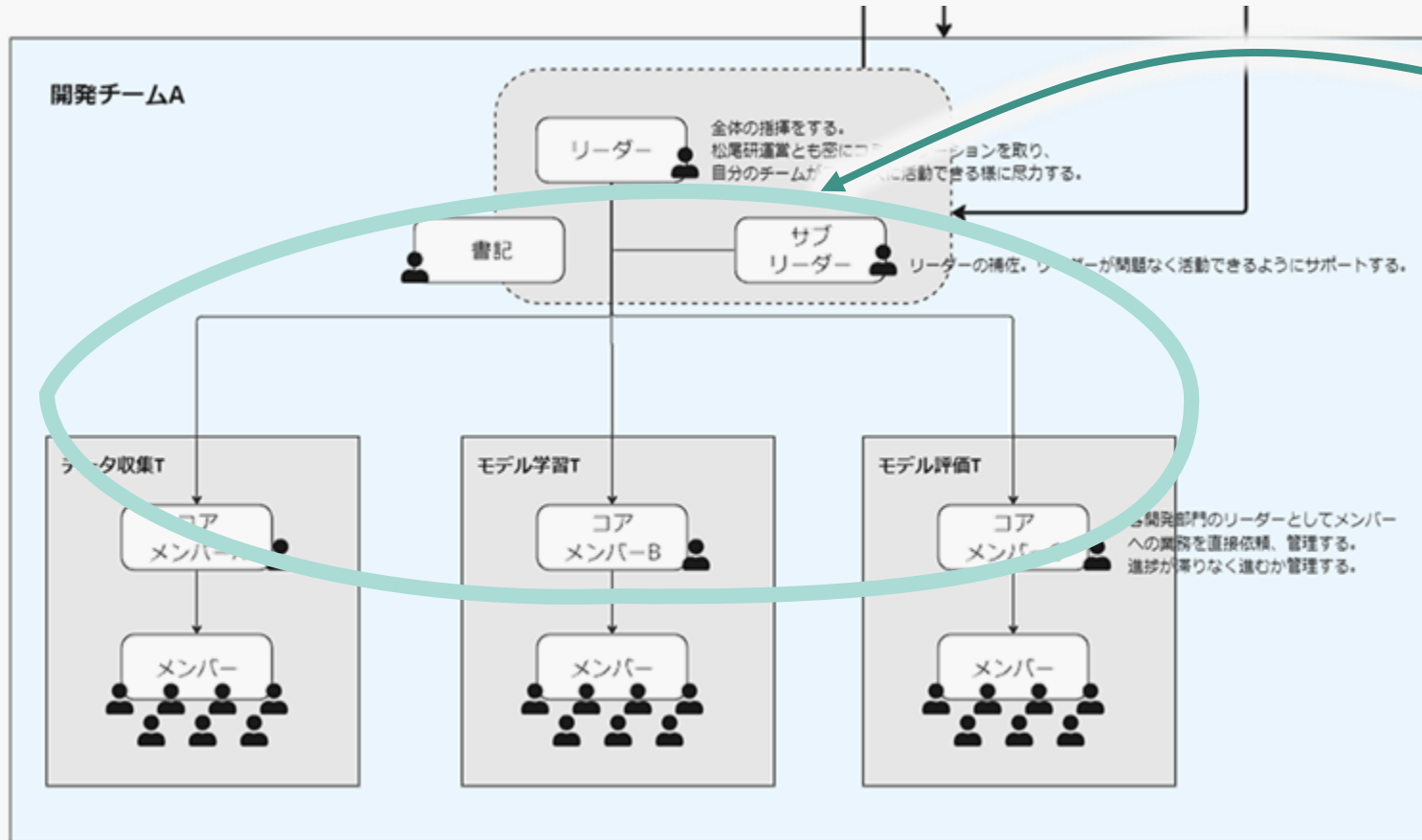
- アイディア出し
- チームの動的な編成（理由は後で述べます）



トップダウンとボトムアップを上手く両立させたい

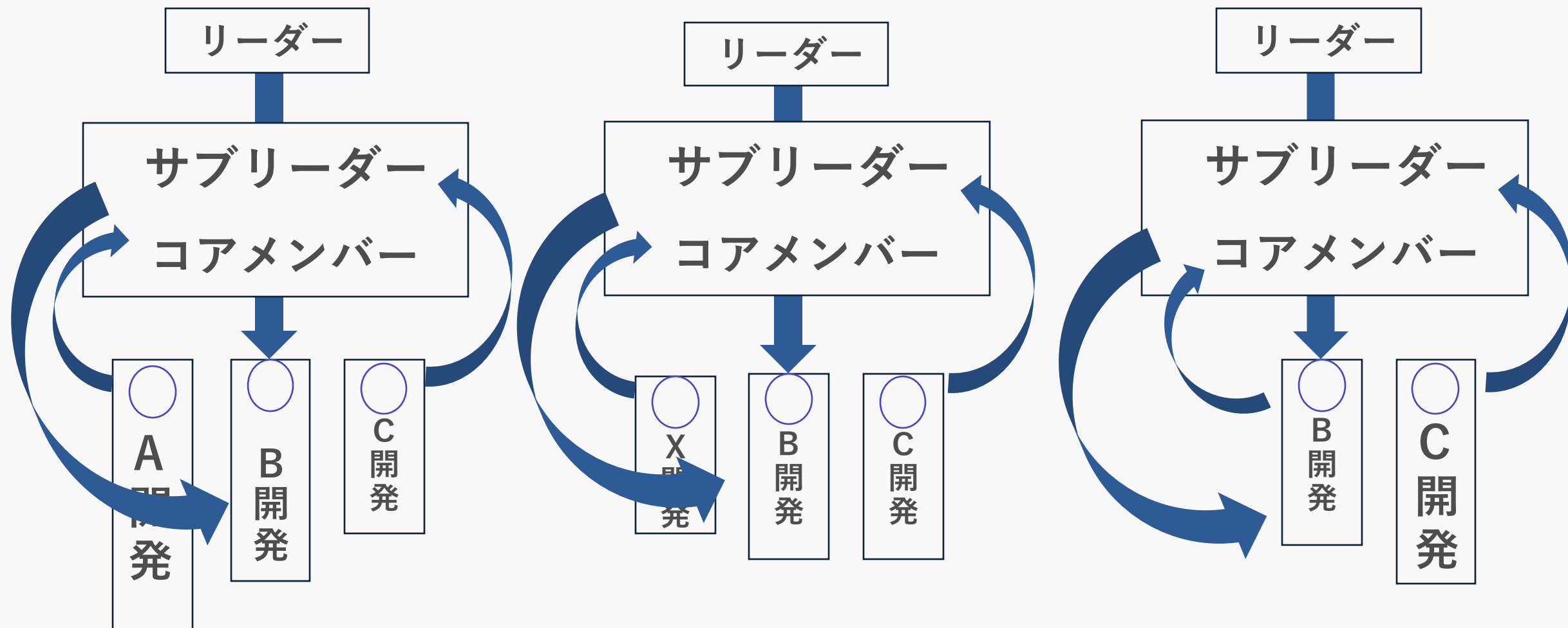


トップダウンとボトムアップを上手く両立させたい

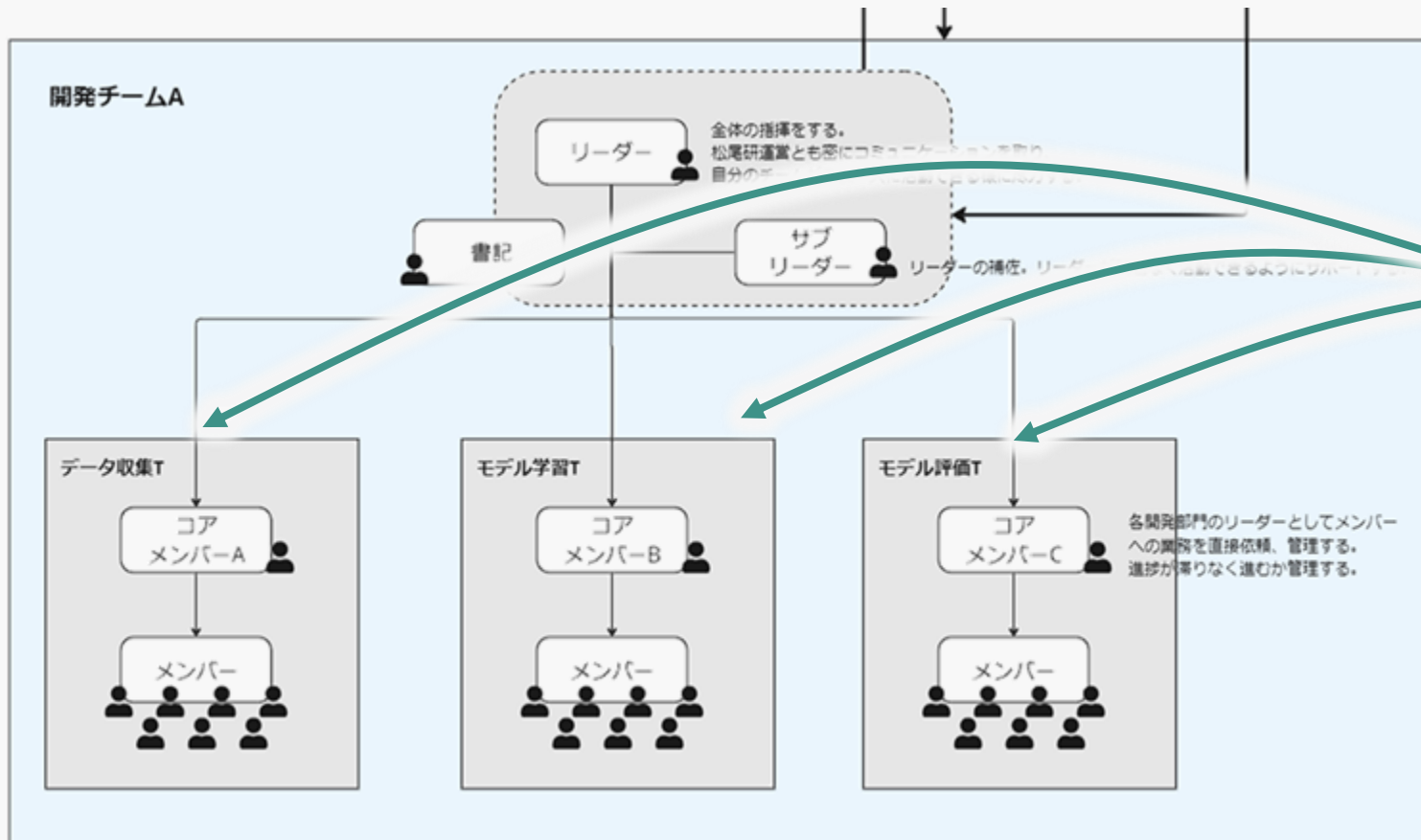


この総合調整機能を大きくしたい

開発の時々に応じたチーム編成を行いたい

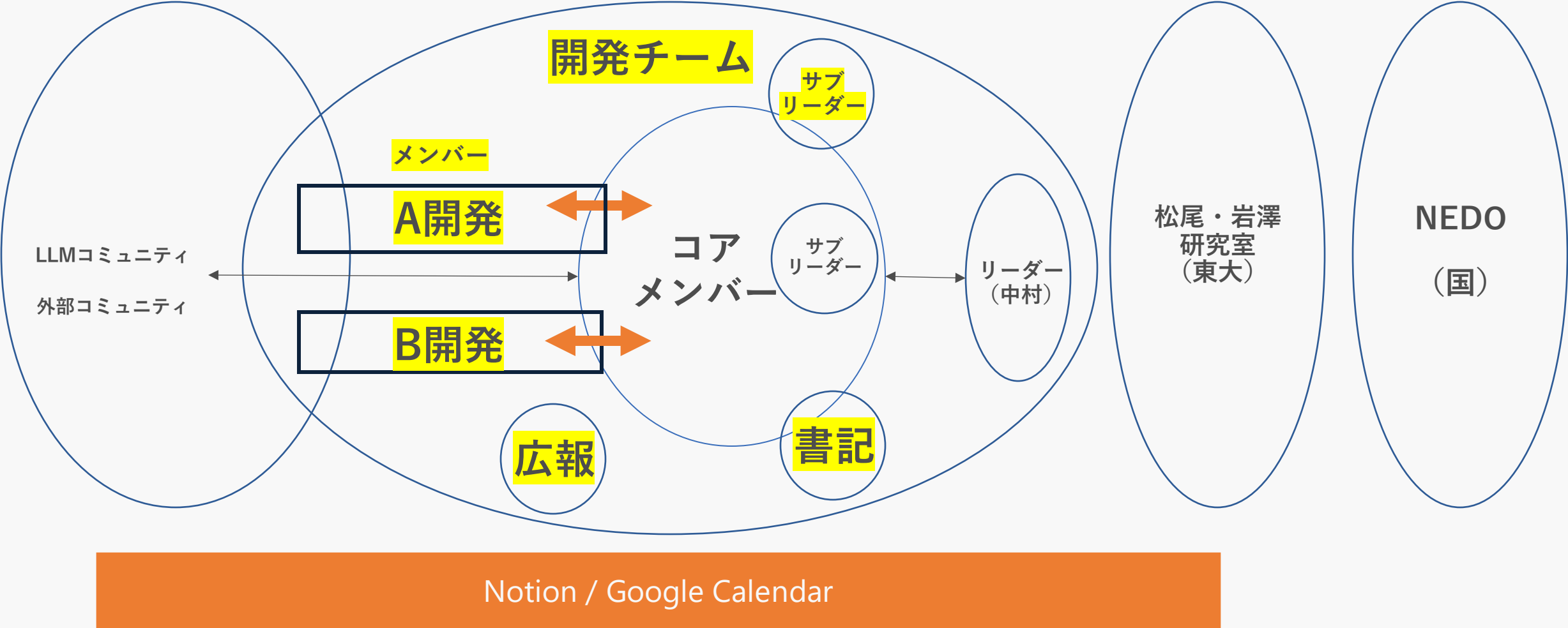


開発の時々に応じたチーム編成を行いたい



メンバーを
必要に応じて割り振る

企画立案・総合調整機能



チームメンバー全員WIN-WINになれるようなチーム作りを進めていきます

- **トップダウンのみでなく、ボトムアップでのチーム運営**
 - トップダウン形式でしっかりプロジェクト管理しつつも、下からの意見をくみ上げる
 - チーム全体で開発
- **匿名の意見フォーム**
 - プロジェクト進行中に何か問題があれば匿名で意見を挙げることが可能
- **Slackを中心とした全体コミュニケーション**
 - 開発について活発なディスカッション
 - 垣根を越えて、みなさんがネットワークを作れるようチーム運営
- **開発についての意見を歓迎**
 - アイディアは常に歓迎
 - 「手を動かしてみたんですが」「サーベイしてみたらこんなアイディアがありました」 歓迎

インセンティブ

本チームに加入することで期待できるメリット

コアメンバー・調整役・貢献者のインセンティブ

- 成果が出れば、共著で論文を投稿
- Note、GitHubを用いたチーム内からの情報公開に際し、お名前を記事に掲載
 - 書き手を募集
 - コミュニティメンバーの方についても記載する

アウトプットを積極的に行い、GENIACのメリットの一つである
「開発者としての社会的認知度向上が期待できる」を最大化

スケジュール

本プロジェクトの想定スケジュール

スケジュール

● 3月末まで

- LLM講座、標準コードを理解する
- 先行事例についての理解を深める
- データセットについての情報収集を行う
 - 可能なら、作り始めたい
- コミュニティメンバーの方も含め、メンバーからのアイデアをまとめる

● 4月以降

- データセット最優先
- モデル改善・環境の整備を行う
- 可能な限り、実行。リソースを無駄にしない

リーダー・
コアメンバーからのメッセージ

最終的なゴールは、このプロジェクトを良い学びとすること

● 今回のLLM開発の大目的

- 500億パラメータの国産LLMを開発すると同時に次世代の国産LLM開発エンジニアを育成する



● 今回のLLM開発を通じてお互い学び合い、次のキャリアやプロジェクトに生かせる状態（もちろん1位にもなりたいけど！）

- みんなでアイデアを出し合ったり、知識を共有する（なので心理的安全性が改めて大事！）
- 新しい考えや手法等の意見に対して、基本前向きにとらえる！（実際にやるかは別として！）
- たくさん経験を得るためにも、いいものや面白そうなアイデアには積極的に挑戦していく！（事前学習できる経験なんて、コストやgpu環境的にそうそうない）
- 他のチームのslack等や論文等も参考にしながらいいものを積極的に組織学習する！
- 上記にあたって、うちのチームもいいものは積極的に発信していく！（情報に関してオープンなスタンス！うちだけでなくみんなでもいいものをつくっていく！）

チーム全体で作っていく，貢献を感じられるようにしたい

開発チーム全員で，さらにコミュニティメンバーも巻き込んで皆で面白いアイデアをLLMとして実装していけたらと思います

- 活発に発言していただける方
- 主体的に活動できる方
- 大きなメンバーを巻き込んで成果を上げていきたい方

このようなチーム活動に対してのマインドを持つ方はぜひ一緒にやりましょう！

このチームに合いそうな方

- **新しいアイデア・ユニークな開発を試したい！！**
 - 使用してみたいデータセットがある！
- ある程度の**裁量権**をもって開発を行いたい！
- **カジュアル**にメンバーと相談し、開発を行いたい！
- Notion等を用いて**プロジェクト管理**をしたことがある！
- **稼働時間が少なくても**、LLM開発がしたい！

ご検討のほどよろしく申し上げます！
改善アイデアは常に募集中です！