2024/02/12 @松尾・岩澤研究室

GENIAC OSS LLM コアメンバー向けプレゼン

中村 仁

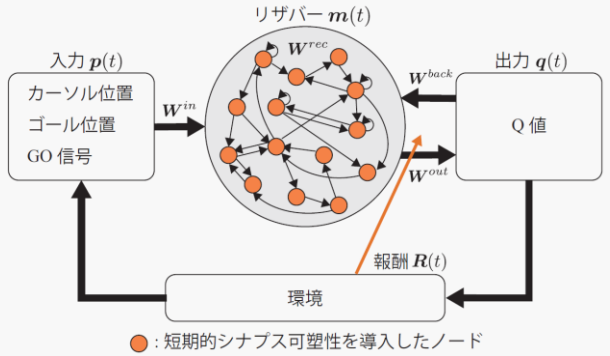
公立はこだて未来大学 システム情報科学部 B4

自己紹介

- **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**リザバー計算**に基づく行動計画の数理モデルと、**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」

LLM開発について

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ィアを多く試したい
  - 面白いアイディアが見つかれば、
    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

ぜひ対処したい領域

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

- 「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。独自制作データも検討中。

「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のDS**
  - ヒンディ：インドに精通した官公庁関連の方々（メンバー予定者）と連携して調査中（採用未定）
- **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
- **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）**

チーミング

チーム全体

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
- 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
- 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

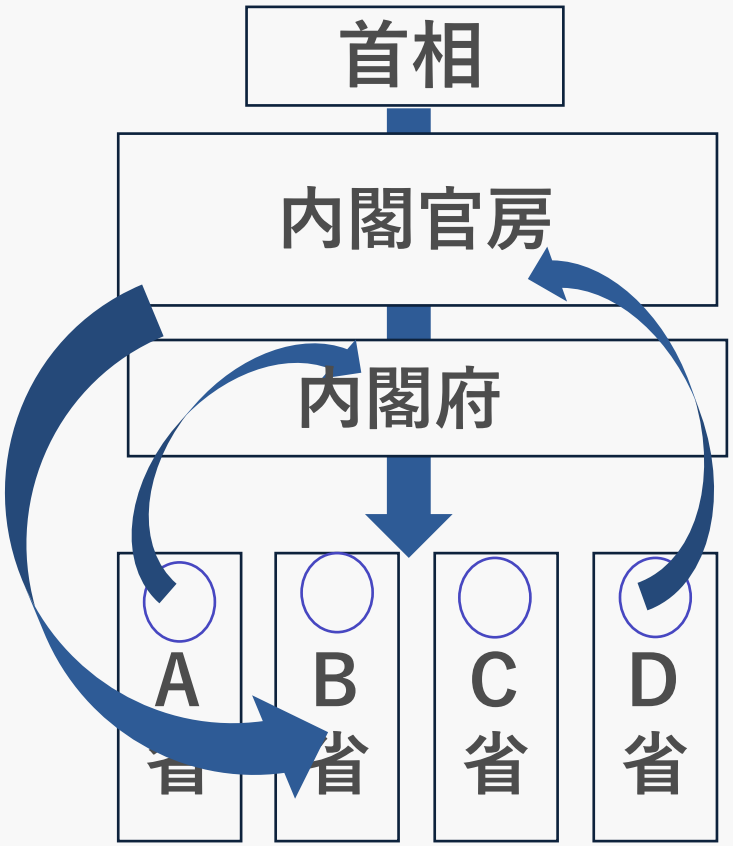
可能なら

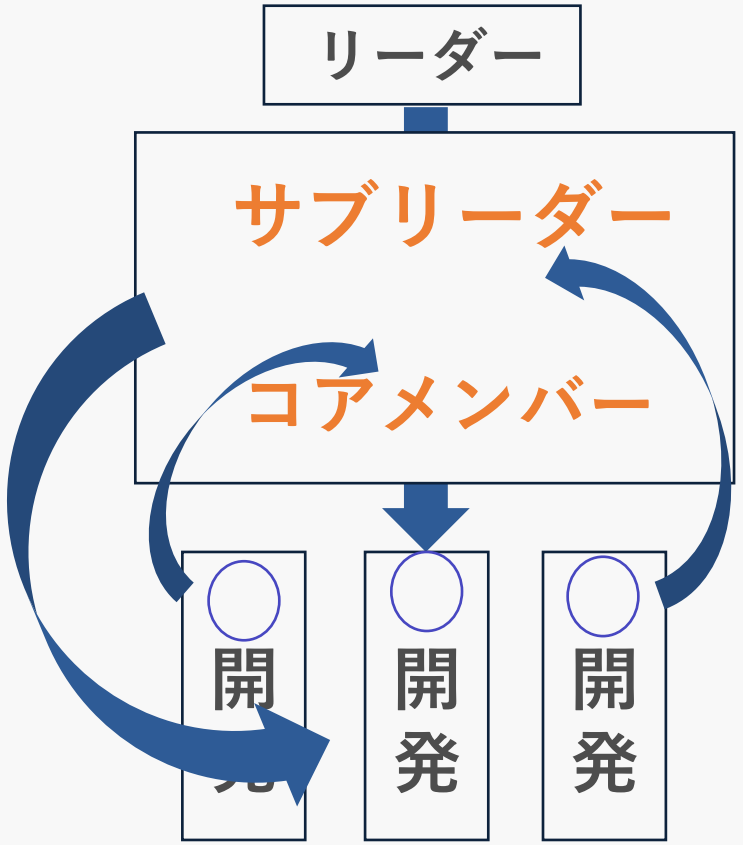
- データセット作り経験
- モデル構造への理解
- 各種アイディア

想定しているプロジェクト管理ツール

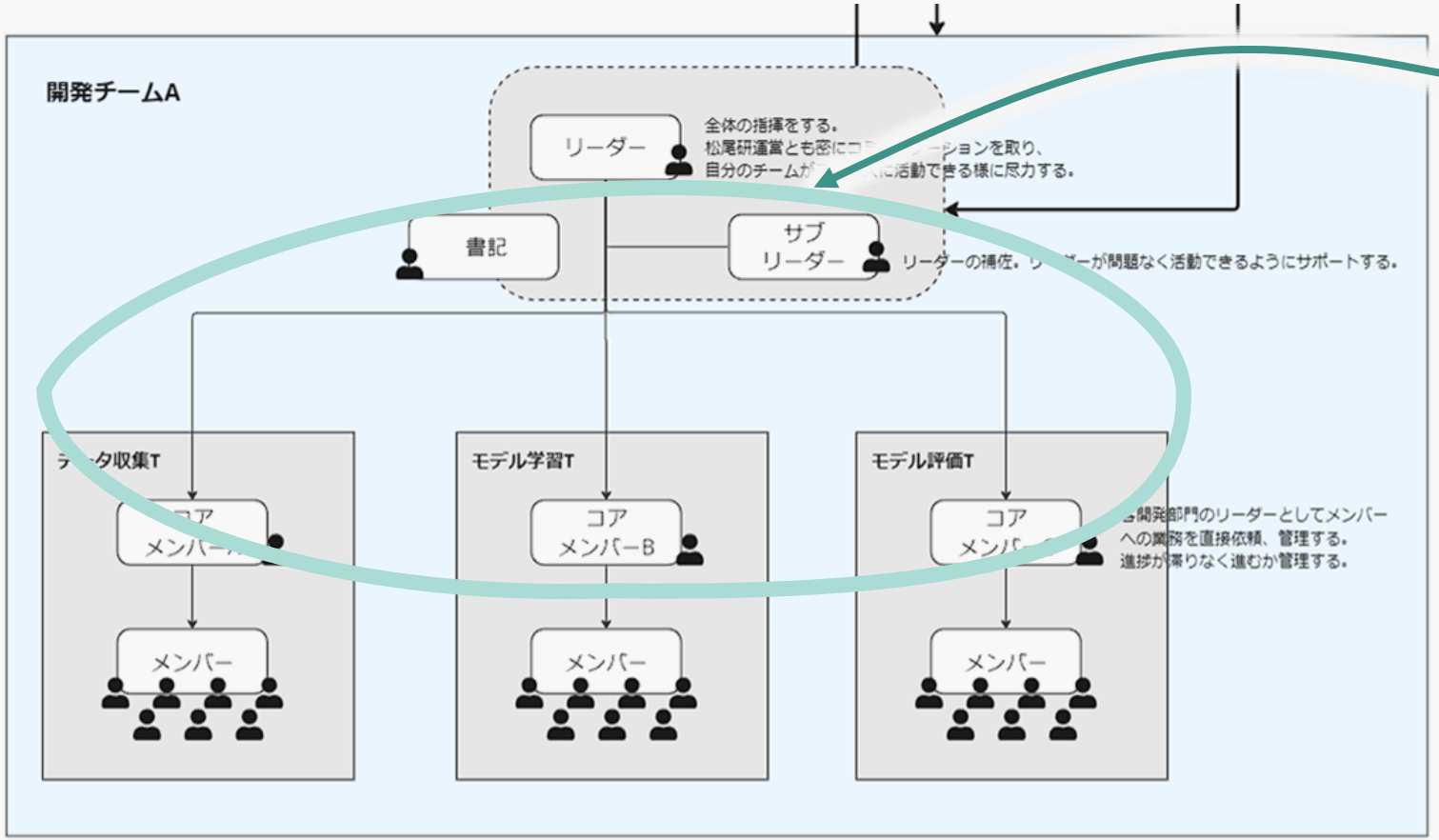
- **Notion**
  - メンバー情報・進捗管理・報告
  - 情報が失われない様、**2つのデータベースを工夫して運用**
- **Google Calendar (Notion Calendar)**
  - 日程調整
- **Slack**
  - 報告連絡相談・コミュニケーション
  - カジュアルなお話は、DMグループで行う予定
- **Miro**
  - アイディア出し
  - **チームの動的な編成**（理由は後で述べます）

トップダウンとボトムアップを上手く両立させたい



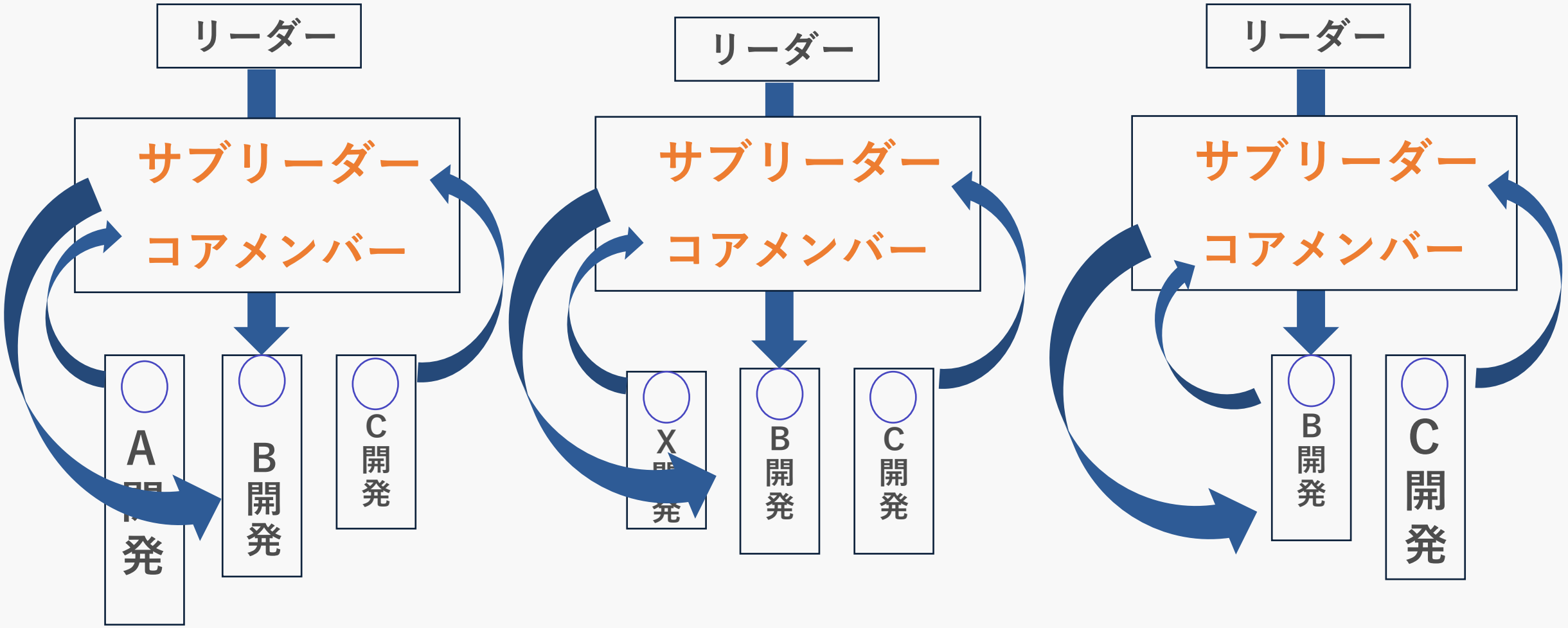


トップダウンとボトムアップを上手く両立させたい

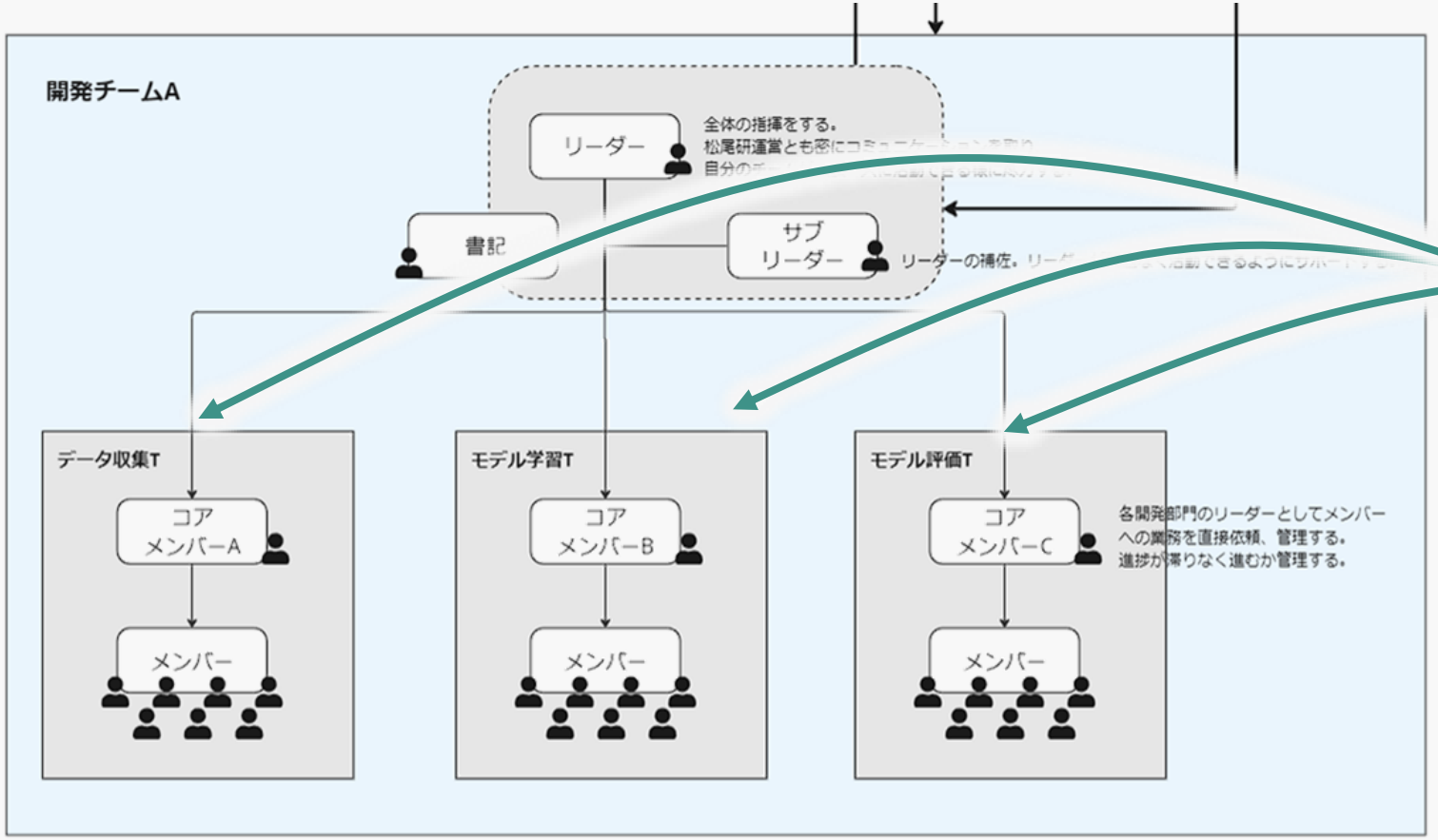


この総合調整機能を
大きくしたい

開発の時々に応じたチーム編成を行いたい

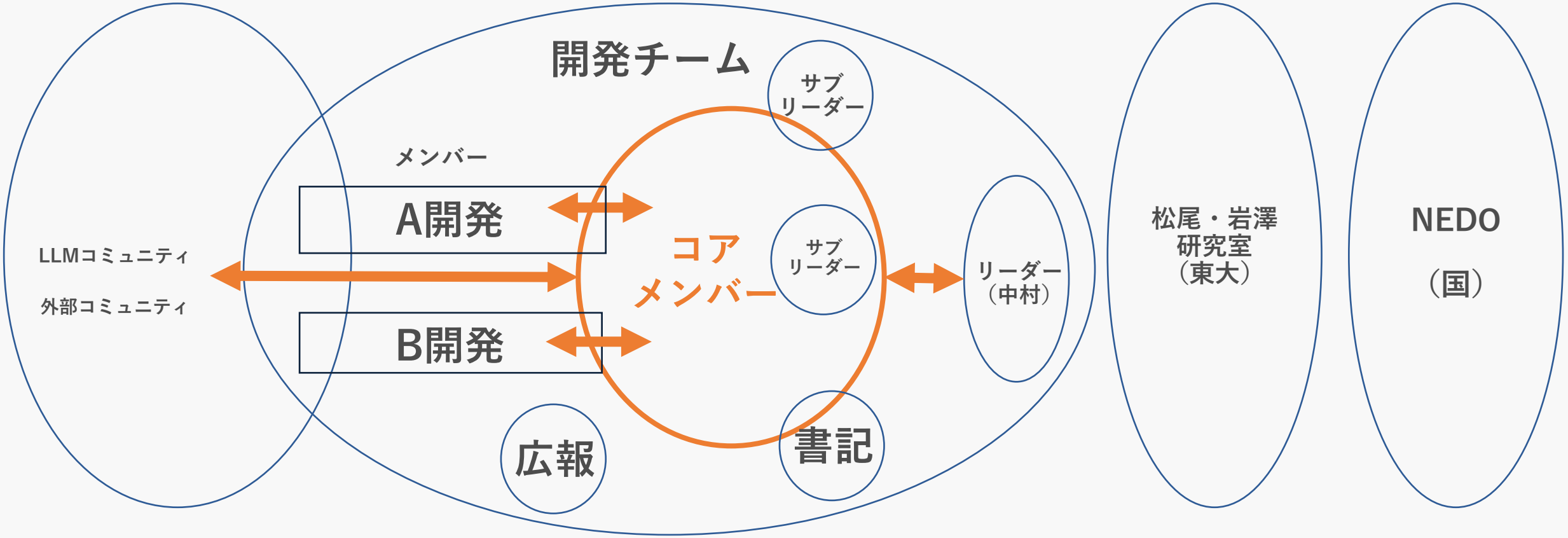


開発の時々に応じたチーム編成を行いたい



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



Notion / Google Calendar

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

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ィアをまとめる
- **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。

このチームに合いそうな方

- **新しいアイディア・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お願いします！
改善アイディアは常に募集中です！