ヒューマンウェアインターンシップ報告書

インターンシップ体験記　（海外インターンシップの場合は英語で記入）

インターンシップを決めるまで

博士後期課程 1 年の夏に投稿していた国際会議論文が採択され、インターン先を本格的に探し始めた。国際会議や学術雑誌などに研究発表することが可能な企業を探し、オムロンサイニックエックスのインターンシップ公募に応募した。CEO である諏訪さんは HW のアドバイザリ委員の一人でもありインターンシップにも興味があった。履歴書を提出した後に一度研究内容についてインタビューがあり、その後採用いただいた。研究内容はいくつか提示いただいた中から、メンターの方々とご相談し、最終的に推論、検索、回答などの行動を繰り返し行う小型言語モデルのエージェントが研究テーマとなった。専門研究とはやや異なる内容であったため、インターンシップ開始前に関連研究のサーベイを行なった。インターンシップの開始時期は、大学での次の論文投稿、ワークショップ参加などに合わせて時期も柔軟に調整いただけた。

インターンシップ先について

OSX はオムロンの研究の最先端を担っており、経験豊富な 20 人程度の研究者と数人の事務員さんで構成されている。リモートでの勤務が可能であり、メンターの方々も基本はリモートで勤務されていた。コンピュータビジョン、ヒューマンインタラクション、ロボティクスなど様々な分野の研究グループがあり、インターンの修士・博士課程の学生も多数参加して研究を進めており、毎月歓迎と送迎を兼ねたランチパーティを行っていた。研究成果、インターンシップ募集要項は随時ホームページに掲載されており、インターンシップで給与をいただくことができる。インターンシップ中の研究においては、モデルの学習に高性能な GPU が必要となり、産総研の ABCI 8×H200 GPU を使用させていただいた。
インターンシップは基本リモートで実施したが、1 週間だけ出社させていただいた。オフィスは東京文京区の東京大学のすぐ近くにあり、東京ドームや上野公園があるような場所だった。オフィスから電車で 2 駅のところのマンションを会社で借りていただきそこに宿泊した。この対面期間は非常に有意義であり、特に手法の最終的な方向性の決定や改良のための集中議論において、研究が大きく前進した。日常的なリモート業務については、週 1 回の定例 Teams ミーティングに加え、GitHub による常時連絡を通じて実装進捗・実験報告・エラー報告・相談などを随時行った。

インターンシップの目的

本インターンシップの目的は、大学院で培った研究能力をさらに発展させるとともに、実社会に近い企業研究環境において、自身の技術や知見を活かして実践的な課題解決に貢献することであった。また、研究者としての視野を広げるために、専門領域にとらわれず、強化学習や知識蒸留といった未経験の技術領域にも挑戦し、理論と実装の両面から課題に向き合う経験を積むことも重要な目的であった。インターンシップを通じて、最先端の研究組織における共同作業、発・議論スタイルといった普段とは異なる環境での企業ならではの研究文化を理解し、今後の研究者人生において活かせる実践的な能力を習得することを目指した。研究成果の社会実装や製品化に向けた社内外に通用する成果を創出し、論文として投稿することで学術的な貢献を果たすことを目標とした。

研究内容

本インターンシップでは、限られた計算資源下（例：ラップトップやスマートフォン）において、小型言語モデルが自律的に推論・検索・回答といった行動を行う「Agentic RAG」の研究に取り組んだ。近年、DeepResearch などの大規模言語モデルを用いた検索エージェントは、調査やレポート作業において、極めて高い性能を示しているが、その推論には大規模な GPU やインフラが必要であり、日常的なアプリケーションやローカル環境での応用には適さない。特に、スマートフォンやエッジデバイス上では、小型モデルによる代替が不可欠である。しかし、既存の小型モデル強化法である強化学習（RL）と知識蒸留（KD）には、それぞれ学習の不安定性や性能の限界などの課題がある。
本インターンシップではそれらの課題を統合的に解決する手法「Distillation-Guided Policy Optimization（DGPO）」を提案し、安定性と自己推論力を両立させる学習フレームワークを実装した。わずか 5 億パラメータ(0.5B)の小型言語モデルを対象に、平均正答率 EM=0.006 という極端に低い初期性能から学習し、最終的には 55 倍以上の改善（EM=0.329）を達成、さらには一部タスクで大型の 30 億パラメータ(3B)のモデルをも上回る成果を得た。

ヒューマンウェアインターンシップ報告書

インターンシップ体験記　（続き）

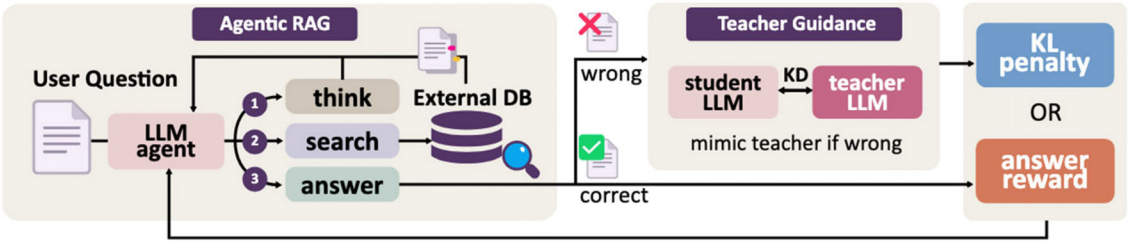

Figure 1: **Distillation-Guided Policy Optimization (DGPO)** establishes a stable reward mechanism by guiding incorrect answers through teacher mimicry.

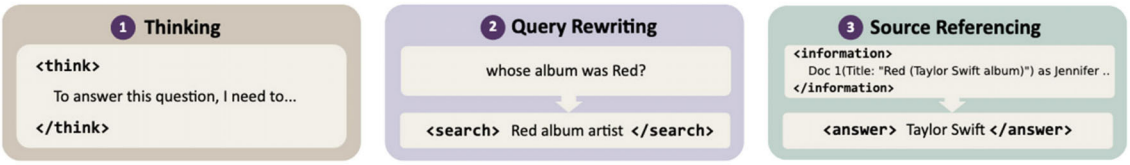

Figure 2: **Agentic RAG Capability (ARC)** characterizes the core capabilities of LLMs required for agentic RAG systems – *thinking*, *query rewriting*, and *source referencing*.

本インターンでは、単なる正答率などの表面的な性能向上にとどまらず、「モデルがどのように考え、検索し、正しく引用するか」といったエージェント的な能力の中身にまで踏み込んだ評価枠組みの構築にも取り組んだ。具体的には、モデルの行動を「①推論」「②検索のためのクエリ生成」「③情報の正確な引用」という三つの要素に分けて定量評価する新しい指標「Agentic RAG Capability（ARC）」を提案し、単純な全体性能のスコアでは見えなかったモデルの詳細な能力を明らかにした。提案する DGPO は、三要素すべてにおいて安定したスコアと高い一貫性を示した。中でも、複数回にわたる再検索が必要な難問に対して、DGPO はその都度クエリを柔軟に生成し直し、最適な探索深度と回答タイミングを見極める動作が見られ、自己判断・自己修正能力に優れた挙動を示した。一方で、DGPO によって検索能力が向上した結果として、検索回数が増加する傾向も確認された。これは、モデルが迷いながらも粘り強く情報探索を続けることの裏返しであるが、実行時間や応答速度といった側面では今後の改善が求められるポイントでもある。このように、評価指標の導入はモデルの強みと課題の両面を明らかにする評価基盤となった。
手法の提案、評価設計と分析を通じて、理論と実験の両面から整理し、自分の研究を深く理解する力を養うことができた。

<u>得られた経験と成長</u>

本プロジェクトは当初、私の専門外である強化学習や蒸留といった領域からスタートしたが、文献調査から理論理解、実装、評価設計まで一貫して主体的に進める中で、技術的な視野が大きく広がった。特に、報酬設計や KL ペナルティの制御といった理論と実装のギャップに苦しみながらも、試行錯誤を重ねることで設計力と問題解決力を鍛えることができた。大学とは異なる企業研究所の環境に触れたことで、出口を見据えた研究姿勢、チームでの意思決定の進め方、非同期的な実装共有と同期的な議論（週次ミーティング＋GitHub での報告）といった普段とは異なる研究スタイルを実体験できた。終盤には東京オフィスに一週間出社し、対面での集中議論を通じて手法設計の完成度を大きく高めることもできた。情報科学分野での研究はリモートのみでも遠隔から効率に実施できるが、もし機会があれば、出社して対面での議論や実際の環境を体験することをおすすめしたい。
本インターンシップでの成果は共同で特許として出願済みであり、現在は国際会議への共著論文投稿と Preprint 公開も進行中である。論文はレビュー中であるため、稼働時間は大幅に減らして今後も継続的に取り組む予定である。小型モデルの限界を押し広げる本研究の意義は大きく、博士人材として、専門外への挑戦、応用指向の視点、研究から発信まで一気通貫で担う経験を通じて、自らの成長を強く実感した。