実験報告

LLMは学生評価のパートナー

石田 亨

香港バプティスト大学客員教授

実験の背景と内容

- 背景：ワークショップコースにおける総合的評価の重要性が増す一方で、教員の負担は大きい。
  なかでも、学生が記述したエッセイの評価は悩ましい。

  効率的で公平な評価の実現が求められている。

- 学部1, 2年生のワークショップ科目を対象に、LLM（ChatGPT 4.0）を用いた評価実験を実施

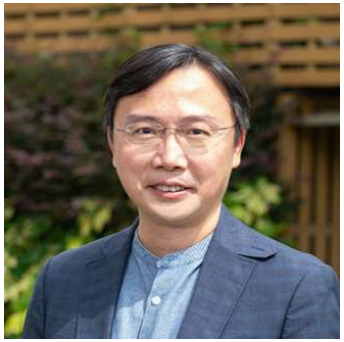
  実験１：複数の教員の評価結果を統合するファシリテーション
  シナリオに基づくケーススタディ

  実験２：幾つかの手法を用いたエッセイの自動評価
  実データを用いた定量的・定性的分析

- 結論：LLMは、教員が構成する評価コミッティのメンバーになりうる能力を示した。

検討チーム（私に加えて）

- Tongxi Liu
  Hong Kong Baptist University　教育学
- Hailong Wang
  早稲田大学/東京大学　土木工学
- Benjamin Luke Moorhouse
  Hong Kong Baptist University　教育学
- William Cheung
  Hong Kong Baptist University　計算機科学



実験の題材としたワークショップ科目

ワークショップ科目

- 早稲田大学創造理工学部の必修のワークショップコース（2021年に実施）
- 3時間ワークショップを7回実施（ideation, interim presentation, prototyping, final presentation）
  - ワークショップは英語で実施され、実施後に提出するエッセイも英語で記述
  - 22名の学生が参加
- エッセイは、6名の教員に5点満点で評価され合算された
  - エッセイは30点満点で、この科目の3割を占める
  - エッセイは概ね良好で、評点は22点から28点の範囲に収まった
- 評価基準は教員間で事前に意見交換はするものの統一されていない
  - 学生の到達点を重視する教員もあれば、学生の成長に注目する教員もいる
  - グループワークを評価する教員もいれば、個性を生かした貢献を評価する教員もいる
  - ライティングを評価する教員もいれば、内省の深さを評価する教員もいる

2020年にコロナ下で実施したワークショップ　https://youtu.be/1v2675Wymu0

評価対象のエッセイ

ワークショップ終了後、学生は以下の記述を含むエッセイの提出を指示された。

- Project description (300-400 words)
  - Title, goal and conclusion of the project
  - Process of the project: how you applied design thinking methods
- Contribution (300-400 words)
  - Your role in the project
  - Your contribution to the project
- Reflection (200-300 words)
  - Write freely but we are interested in what you learned from the series of workshops not only professional knowledge and skills but also how to contribute to and facilitate the workshops

実験1

複数の教員の評価結果を統合

LLMに提示したシナリオ（実体験をデフォルメしたもの）

複数の教員が、ある学生のエッセイについて評価を述べています。
あなたは異なる意見を調整し結論を導くファシリテータです。

- 教員１:『この学生は非常に意欲があり、熱心に取り組んだことがエッセイから伺える。技術的にはよく理解したかどうか分からない面もあるが、意欲を買って成績はAとしたい。』
- 教員２:『この学生は、文章上は熱心さをアピールしているが、実際にチームにどう貢献しかたは明確には書かれていない。また、技術的にもよく理解しているとは思えない。従って成績はBだろう。』
- 教員３:『この学生は、他の学生に比べて、著しく文章の量が少ない。エッセイの分量は上限しか決めていないので、量が少ないという理由で不合格にはできないが、書けないということは、熱心に取り組んでいなかったということではないか。従って、成績はCだろう。』

LLMによる評価の統合

- 意欲の評価
  教員1が指摘するように、学生の意欲や熱心さを評価することは重要です。しかし、教員2と教員3の意見を考慮すると、この意欲や熱心さが実際の成果や貢献にどれだけ繋がっているかは不明確です。
- 技術的理解
  教員2が指摘するように、技術的な理解が不十分である可能性が考えられます。教員1も技術的な理解については確信を持っていないようです。
- エッセイの分量
  教員3が指摘するように、エッセイの分量が少ないことは、学生が十分に取り組んでいなかった可能性を示唆しています。しかし、分量だけで評価するのは公平ではないとも言えます。

結論として、3人の教員の意見を総合すると、成績はBとするのが妥当と思います。

背後にある理論と論文をLLMに尋ねる

結論の背後にある理論と関連論文を示してください。

- 多面評価（360度フィードバック）: 複数の評価者からのフィードバックを集め、総合的な評価を行う手法です。このケースでは、3人の教員からの意見を取り入れて、総合的な判断を下しています。
  Bracken, D. W., Timmreck, C. W., & Church, A. H. (2001). The Handbook of Multisource Feedback. Jossey-Bass.
- 公平性の原則: 評価の公平性は、教育の評価において非常に重要な要素です。結論では、エッセイの分量だけでの評価は公平ではないとの観点から、総合的な評価を行っています。
  Colquitt, J. A. (2001). On the dimensionality of organizational justice. Journal of Applied Psychology, 86(3), 386.
- 成果主義の評価: 意欲や努力だけでなく、実際の成果や貢献も評価の要素として考慮するべきです。教員2の意見に基づき、実際の貢献や技術的な理解を評価の要素として取り入れています。
  Lawler III, E. E. (2003). Reward practices and performance management system effectiveness. Organizational Dynamics, 32(4), 396-404.

これらの理論を採用した理由は、多角的な視点からのフィードバック、公平性、そして実際の成果を重視することが、よりバランスの取れた評価を行うために重要であると考えられるからです。

後続するシナリオ（結構長いです）

一人の教員が遅れて以下の評価を提出してきました。

『この学生のワークショップでの活動を観察してきた。当初は自己中心的なところがあったが、徐々に共同作業ができるようになった。その進歩を評価したい。到達点としてはAとは言えないが、成長を評価するなら明らかにAと思われる。』

さらに以下の指摘がありました。

『複数の学生から、この学生はチームの協調を乱すことが多かったと報告されています。』

この件が議論になると、教員の意見は割れました。ある教員は、学生の相互評価の重要性を指摘し、ある教員は、そもそもそういう評価を学生に依頼していないと主張しました。

成績評価の議論の途中で、経験豊かな教員が参加してきました。本来この教員は他の教員から一目置かれています。

『彼には才能がある。彼の一言でチームの議論が変わったよ。ワークショップ科目は、通常科目とは違う。こういう学生を評価できないなら、そもそもこんな科目はいらないよ。』

最後にLLMに採点基準の作成を依頼（教育学の専門家が感心した）

このシナリオを踏まえて、エッセイの採点基準と、背後にある理論・論文を示してください。

1. **内容の理解と技術的知識** (30%):
   - エッセイ内での技術的な内容の正確さと理解の深さを評価します。
   - 使用された技術やツールに関する知識や理解を示す部分を評価します。
2. **意欲と熱心さ** (20%):
   - 学生のワークショップへの取り組みや学習意欲を示す部分を評価します。
   - エッセイの中での自らの学びや成長を示す記述を評価します。
3. **チームワークと協力** (20%):
   - チーム内での役割や貢献、協力の様子を示す部分を評価します。
   - 他の学生との協力やコミュニケーションの取り組みを示す記述を評価します。
4. **特定の能力や貢献** (15%):
   - エッセイ内での特定の能力や才能、特にチームの議論や方向性に影響を与えた部分を評価。
   - 他の学生や教員からのフィードバックや評価を参考にします。
5. **実践的学習と経験** (15%):
   - ワークショップ中の実際の経験や学びを示す部分を評価します。
   - 実際の問題解決やプロジェクトの取り組みを示す記述を評価します。

引用された理論と論文（非常に勉強になりました）

この実験のログ（英文）
が必要な方は
toru.Ishida@gmail.com
までご連絡下さい。

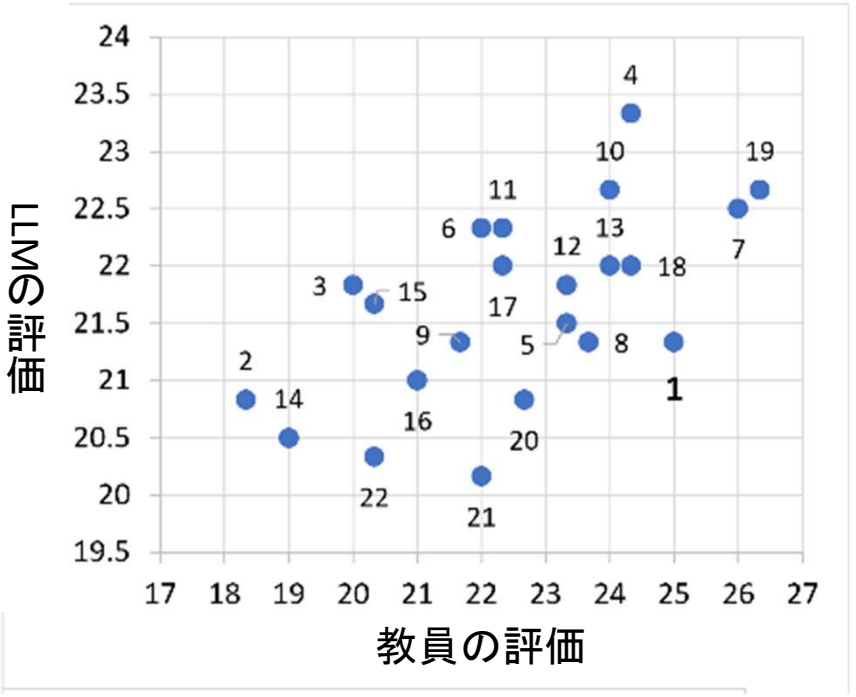
- ピア評価 (Peer Assessment)
  Topping, K. J. (1998). Peer assessment between students in colleges and universities. Review of Educational Research, 68(3), 249-276.
- 中央値を用いた評価
  Wilcox, R. R. (2012). Introduction to robust estimation and hypothesis testing. Academic Press.
- 評価の公平性と透明性
  Stiggins, R. J., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. Theory into Practice, 44(1), 11-18.
- 成長思考セット (Growth Mindset)
  Dweck, C. S. (2006). Mindset: The new psychology of success. Random House Incorporated.
- タレントマネジメント (Talent Management)
  Collings, D. G., & Mellahi, K. (2009). Strategic talent management. Human Resource Management Review, 19(4), 304-313.
- 形成的評価 (Formative Assessment)
  Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7-74.
- 多様性の尊重 (Respect for Diversity)
  Banks, J. A. (2004). Diversity and citizenship education: Global perspectives. Jossey-Bass.
- 実践的学習 (Experiential Learning)
  Kolb, D. A. (1984). Experiential learning: Experience as the source of learning and development. Prentice-Hall.
- 合意形成の理論 (Consensus Building)
  Innes, J. E., & Booher, D. E. (2004). Reframing public participation: strategies for the 21st century. Planning theory & practice, 5(4), 419-436.

実験2

エッセイの自動評価

LLMに評価を任せる

LLMは自ら評価基準となるルーブリックを生成し、それに基づいて評価を行う。LLMが持つランダム性のため、評価の度に異なるルーブリックが生成される。



LLMに与えるデータは以下のとおり

1. ワークショップの説明（概ね500語）
2. 学生のエッセイ：Project description, Contribution, Reflectionからなるエッセイ（800語から1100語）
3. 評価プロセス：ワークショップの説明を基にルーブリックを生成し、背景となる理論を説明。
   そのルーブリックを用いてエッセイを30点満点で評価し、評点の根拠を説明。

22名のエッセイについて、それぞれ6回の評価を行い、その平均値を得点とした。
LLMと教員の評価値の相関係数は0.611。

教員がルーブリックを作成

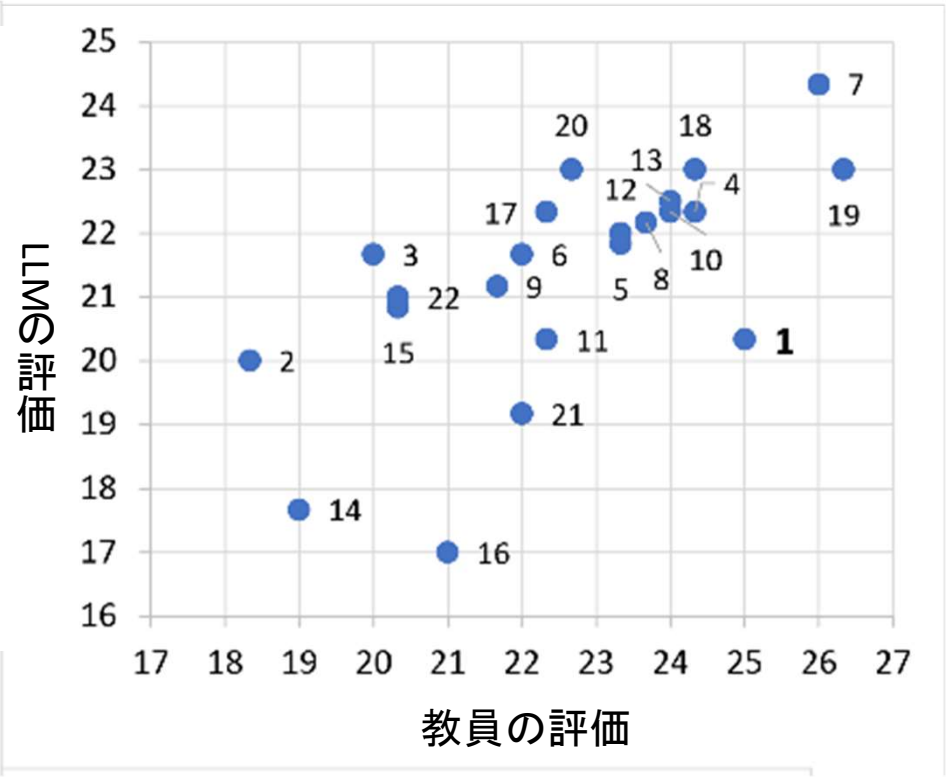
1. Technical Knowledge and Application (10 Points)
- Understanding of Concepts: The student's grasp of technological and theoretical concepts relevant to the project.
- Practical Application: Effectiveness in applying technical knowledge in practical situations.
- Innovation and Problem Solving: Creativity and innovation in addressing challenges and proposing solutions.

2. Teamwork and Collaborative Skills (10 Points)
- Individual Role and Contribution: The student's clarity in defining and fulfilling their role, contributing to the project.
- Team Interaction and Communication: Ability to communicate and collaborate effectively within the team.
- Peer Engagement: Participation in peer learning, support to team members, and contribution to team dynamics.

3. Reflective Learning and Personal Growth (10 Points)
- Self-Reflection and Insights: Depth of self-reflection on learning and development throughout the project.
- Design Thinking and Process: Application of design methods and management of the project process.
- Skill and Attitude Development: Growth in soft skills, such as critical thinking, adaptability, and communication.

教員が作成したルーブリックで評価

ルーブリックを指定し、 LLMにエッセイの評価を依頼。全ての評価において共通のルーブリックが適用される。



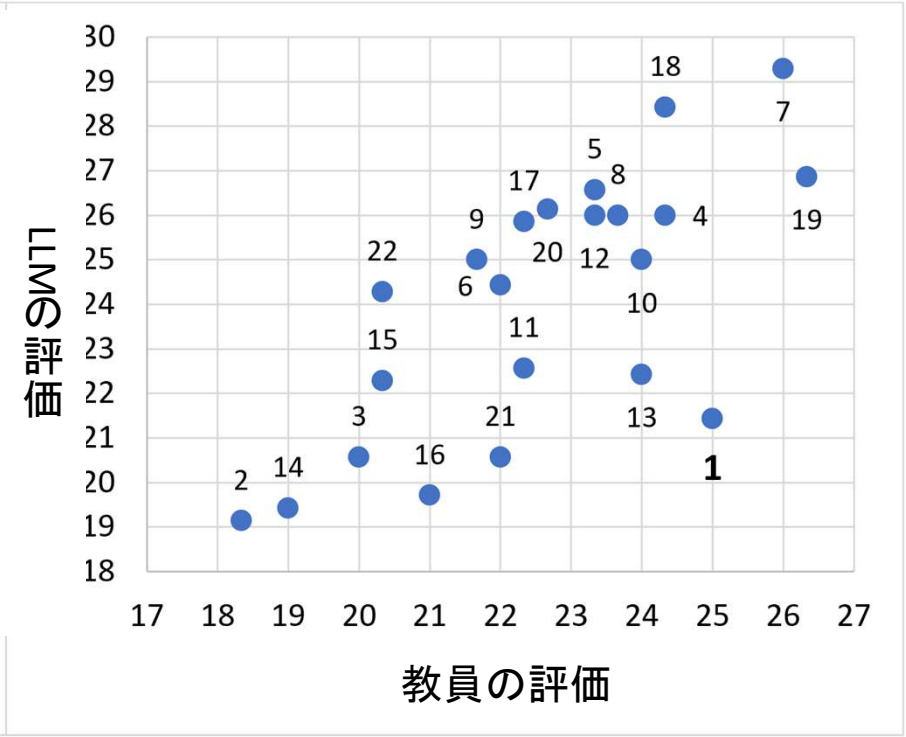
LLMに与えるデータは以下のとおり。

1. ワークショップの説明（概ね500語）
2. 学生のエッセイ（800語から1100語）
3. 教員が作成したルーブリック
4. 評価プロセス：指定したルーブリックを用いて学生のエッセイを30点満点で評価し、なぜその評点となったのか説明。

22名のエッセイについて、それぞれ6回の評価を行い、その平均値を得点とした。
LLMと教員の評価値の相関係数は0.657。

一対比較による評価

LLMはエッセイを個々に評価するため、全体を見渡してエッセイの順位付けをしていない。そこで、エッセイの一対比較をLLMに依頼。



LLMに与えるデータは以下のとおり。

1. ワークショップの説明（概ね500語）
2. 学生のエッセイ（800語から1100語）
3. 教員が作成したルーブリック
4. 評価プロセス：22名の学生のエッセイの総当たり戦を行う。

一対比較では勝者と敗者が明確になる傾向がある。一対比較のコメントは、学生へのフィードバックには適切ではない。フィードバックには個別評価を用い、成績はエッセイ間の比較評価を経て決定するのが適切。LLMと教員の評価値の相関係数は0.716。

LLMの評価レポートを定性的に分析

1.　LLMは教員に比肩する評価能力を備えている

LLMは教育学の専門的な知識を背景に多様なルーブリックを作り出すことができる。また、LLMは、指定されたルーブリックに基づいてエッセイの評価を行い、適切なコメントを生成できる。

2.　LLMによる評価で観察されるランダム性は錯乱ではなく、それぞれに論理的一貫性を持つ多様性

LLMの評価は実行するたびに結果が変化するが、評価結果を読むとそれぞれに筋が通っている。まるで、LLMの中にいる多様な教員が次々に登場して意見を述べているかのようである。

3.　教員が感銘を受け高く評価したエッセイをLLMは高く評価できない場合がある

LLMは、教員のような教育経験の蓄積がない。このため、「この学生には特殊な能力がある」「今後の成長が期待できる」と判断して高得点を与えることはできない。一方、LLMの評価は教員の過剰な思入れを抑制する効果もある。

LLMは万感の書を読んだが、未だ万里の路を歩んではいない？

まとめ

LLMに以下の３通りでエッセイの評価を依頼

1. 評価を任せる
2. ルーブリックを指定して評価を依頼
3. エッセイの一対比較を依頼。

- 評価結果を定量分析したところ、ルーブリックを指定し一対比較を実施すると、LLMと教員の評価に強い相関が認められた。一方、評価の質や安定性に不安が残った。
- そこで、LLMの評価コメントの定性分析を行い以下を示した。
  1. LLMが教員に比肩する評価能力を持つこと
  2. LLMの評価の振れは錯乱ではなく多様性と解すべきであること
  3. 人間とLLMの評価は異なる場合があり相補的でありうること
- LLMは教員のアシスタントではなく、教員チームの一員としてエッセイ評価に参加するだけの能力を有している。教員が不足する現場では、教員とLLMが評価を行い結果を突き合わせることが有効である。（LLMの評価も学生に開示する？）

ご清聴ありがとうございました