

CS 2731

Introduction to Natural Language Processing

Session 14: Project proposal presentations

October 16, 2024

Schedule

1. Geonyeong, Kiran, Hugh, Carolina
2. Maanya, Jerry, Alex
3. Yushui, Yifang, Zhuochun
4. John, Rojin, Xianglong
5. Yansheng, Xiaoyan, Shijai
6. Dilip, Anveshika, Akshat
7. Joel, Jiyang

Instructions

- Plan for **7 min presentations max** not including Q&A
- Cover at least these key points
 - Project motivation (what is the value of this work?)
 - Super briefly, what 1-2 other related papers have done
 - What data you are planning to use
 - What approach/methods will you be taking
 - Evaluation of your approach (or dataset, if it's a dataset contribution)
- Put your slides in this presentation after your project name slide by **class session, 2:30pm on Wed Oct 16**

1. Geonyeong, Kiran, Hugh, Carolina

Automated Extraction of Cellular Niches from Scientific Literature

10/15/2024

Hugh Galloway (hug18@pitt.edu),
Kiran Shridhar Alva (kiranshridhar@pitt.edu),
Geonyeong Choi (gec108@pitt.edu)

Overview

- The goal of this project is to compile a large set of papers which analyze spatially resolved single-cell data
- After the creation of this dataset we want to automatically extract cellular neighborhoods (CNs), also referred to as 'niches' or 'recurrent cellular neighborhoods'
- CNs are recurring patterns of cells that group together in tissue and are can be used to predict important clinical outcomes like patient response to therapy
- There are many papers which measure single-cell gene expression and colocalization of cell types but to date no one has compared findings across these papers. Our automated meta-analysis would allow us to immediately understand trends in the results across this incredibly hot field.

Project Deliverables

- The project deliverables are the following:
 1. Extract a dataset of relevant spatial single-cell papers which contain cellular niches (search for papers in biorxiv/PubMed, extract using abstract and title, extract the papers as pdfs)
 2. Develop a method for extracting cellular neighborhoods and their associated biology from the papers in an automated manner (evaluate performance on a small hand-labeled subset)
 3. Run the aforementioned method on all papers in our extracted collection and create a table or database that matches each paper to its cellular niches, associated biology and important data like disease types and tissue of origin

Single-cell spatial immune landscapes of primary and metastatic brain tumours


<https://doi.org/10.1038/s41586-022-05680-3>

Received: 24 March 2022

Accepted: 22 December 2022

Published online: 1 February 2023

Open access

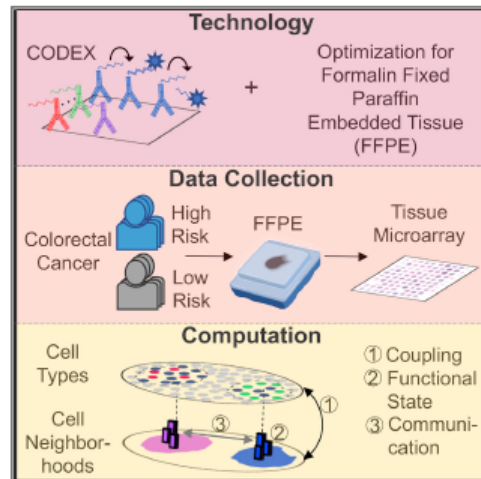
 Check for updates

Elham Karimi^{1,20}, Miranda W. Yu^{1,2,20}, Sarah M. Maritan^{1,3,20}, Lucas J. M. Perus^{1,2,20}, Morteza Rezaejad⁴, Mark Sorin^{1,5}, Matthew Dankner^{1,3}, Parvaneh Fallah⁶, Samuel Doré^{1,5}, Dongmei Zuo¹, Benoit Fiset¹, Daan J. Kloosterman⁷, LeeAnn Ramsay¹, Yuhong Wei¹, Stephanie Lam⁸, Roa Alsajjan^{9,10}, Ian R. Watson^{1,11}, Gloria Roldan Urgoiti^{12,13}, Morag Park^{1,6,11}, Dieta Brandsma¹⁴, Donna L. Senger^{6,15}, Jennifer A. Chan^{1,3,16}, Leila Akkari⁷, Kevin Petrecca^{10,17}, Marie-Christine Guiot^{10,17,18}, Peter M. Siegel^{1,3,11,19}, Daniela F. Quail^{1,2,3} & Logan A. Walsh^{1,5}✉

Single-cell technologies have enabled the characterization of the tumour microenvironment at unprecedented depth and have revealed vast **cellular diversity among tumour cells and their niche**. Anti-tumour immunity relies on cell–cell relationships within the tumour microenvironment^{1,2}, yet many single-cell studies lack spatial context and rely on dissociated tissues³. Here we applied imaging mass cytometry to characterize the immunological landscape of 139 high-grade glioma and 46 brain metastasis tumours from patients. Single-cell analysis of more than 1.1 million cells across 389 high-dimensional histopathology images enabled the spatial resolution of immune lineages and activation states, revealing differences in immune landscapes between primary tumours and brain metastases from diverse solid cancers. **These analyses revealed cellular neighbourhoods associated with survival in patients with glioblastoma**, which we leveraged to identify a unique population of myeloperoxidase (MPO)-positive macrophages associated with long-term survival. Our findings provide insight into the biology of primary and metastatic brain tumours, reinforcing the value of integrating spatial resolution to single-cell datasets to dissect the microenvironmental contexture of cancer.

Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front

Graphical Abstract



Authors

Christian M. Schürch, Salil S. Bhatt, Graham L. Barlow, ..., Nikolay Samusik, Yury Goltsev, Gary P. Nolan

Correspondence

christian.m.schuerch@gmail.com (C.M.S.),
gnolan@stanford.edu (G.P.N.)

In Brief

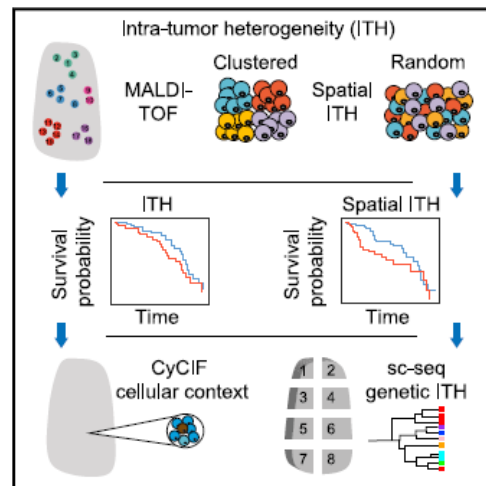
A multiplexed tissue imaging and computational analysis framework applied to colorectal cancer allows interrogation of how spatial organization of the immune tumor microenvironment is linked to clinical outcomes.

Highlights

- FFPE-CODEX multiplexed tissue imaging of 56 markers in 140 tissues of 35 CRC patients
- Cellular neighborhoods reveal spatial organization of the tumor microenvironment
- Altered organization of tumor and immune components in low- versus high-risk patients
- Local enrichment of PD-1⁺CD4⁺ T cells correlates with survival in high-risk patients

Spatial intra-tumor heterogeneity is associated with survival of lung adenocarcinoma patients

Graphical abstract



Authors

Hua-Jun Wu, Daniel Temko, Zoltan Maliga, ..., Nicholas Navin, Robert J. Downey, Franziska Michor

Correspondence

downey@mskcc.org (R.J.D.), michor@jimmy.harvard.edu (F.M.)

In brief

Using three orthogonal spatially resolved molecular profiling technologies, Wu et al. identified two groups of lung adenocarcinomas with distinct patterns of geographic diversification. They found that these two groups differed in their survival outcomes and found evidence suggesting that the observed patterns may be linked to differences in tumor cell motility.

Highlights

- Lung adenocarcinomas show “random” or “clustered” GD
- Random GD tumors have significantly poorer survival
- Random GD tumors are characterized by decreased cell adhesion
- Clustered GD tumors have high levels of tumor cell-interacting endothelial cells

RESEARCH ARTICLE

The Spatial Landscape of Progression and Immunoediting in Primary Melanoma at Single-Cell Resolution

Ajit J. Nirmal^{1,2,3}, Zoltan Maliga^{1,2}, Tuulia Vallius^{1,2}, Brian Quattrochi⁴, Alyce A. Chen^{1,2}, Connor A. Jacobson^{1,2}, Roxanne J. Pelletier^{1,2}, Clarence Yapp^{1,2}, Raquel Arias-Camison^{1,2,4}, Yu-An Chen^{1,2}, Christine G. Lian⁴, George F. Murphy⁴, Sandro Santagata^{1,2,4}, and Peter K. Sorger^{1,2,5}

ABSTRACT

Cutaneous melanoma is a highly immunogenic malignancy that is surgically curable at early stages but life-threatening when metastatic. Here we integrate high-plex imaging, 3D high-resolution microscopy, and spatially resolved microregion transcriptomics to study immune evasion and immunoediting in primary melanoma. We find that recurrent cellular neighborhoods involving tumor, immune, and stromal cells change significantly along a progression axis involving precursor states, melanoma *in situ*, and invasive tumor. Hallmarks of immunosuppression are already detectable in precursor regions. When tumors become locally invasive, a consolidated and spatially restricted suppressive environment forms along the tumor–stromal boundary. This environment is established by cytokine gradients that promote expression of MHC-II and IDO1, and by PD1–PDL1-mediated cell contacts involving macrophages, dendritic cells, and T cells. A few millimeters away, cytotoxic T cells synapse with melanoma cells in fields of tumor regression. Thus, invasion and immunoediting can coexist within a few millimeters of each other in a single specimen.

SIGNIFICANCE: The reorganization of the tumor ecosystem in primary melanoma is an excellent setting in which to study immunoediting and immune evasion. Guided by classic histopathology, spatial profiling of proteins and mRNA reveals recurrent morphologic and molecular features of tumor evolution that involve localized paracrine cytokine signaling and direct cell–cell contact.

Technology Background

- Spatial Transcriptomics is a relatively new and exceedingly popular method for analyzing tissue biology
- Named Nature's method of the year in 2020 [1]
- This technology allows us to pinpoint individual cells within tissue while understanding their types and function
- Basic idea is that now we have gene expression within individual cells as well as the coordinates of those cells
- Cellular neighborhoods are almost always the final deliverable produced with this data collection method.



[1] states that RNA-seq is like a smoothie (gives us broken down components of tissue), single-cell RNA-seq is like a fruit salad (we know what the individual components are now), and spatial transcriptomics is like a fruit tart (we know the individual components and their arrangement)

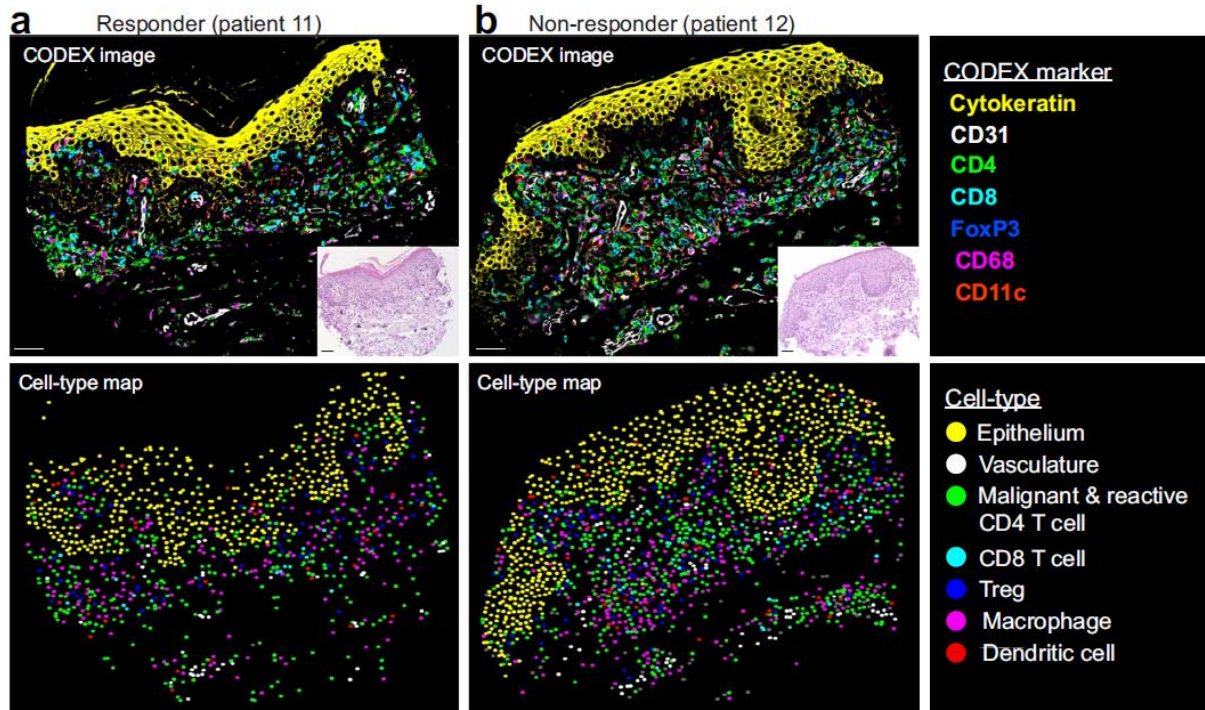
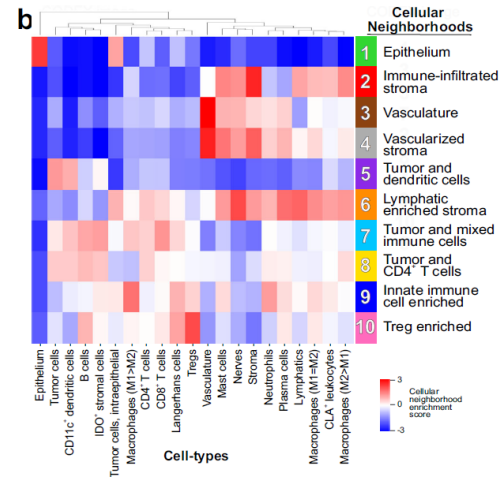


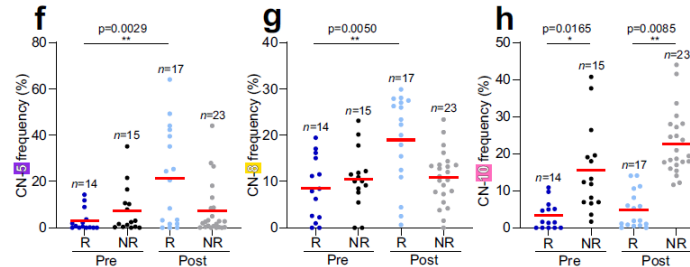
Figure from [2]: **a.)** Shows a raw multiplexed image of a tissue with cell-type markers, below is a map of the typed cells as points. These images are associated with response to immunotherapy, note the interaction between immune cells and tumor cells (epithelium). **B.)** Shows raw image and cell-type map from non-responder, note the lack of interaction between immune cells and tumor cells.

Relevance

- The goal of spatially resolved single-cell analysis is to understand the tumor microenvironment
- Understanding of the tumor microenvironment is critical as researchers push to develop immunotherapies for cancer
- The goal of immunotherapy is to train the immune system to attack cancer cells, thus reducing side effects from treatment and improving odds of lasting response
- It's worth noting that there are many trends in this analysis of CN's, examples include
 - Immune cell and tumor cell interaction tends to correlate with treatment response and survival
 - Interaction between immune cells and immune regulators like Tregs and macrophages tends to be associated with poor response and survival
- Thus, tracking these trends across different data collection platforms, disease types, treatment types and tissue types is very interesting



Heatmap showing identified cellular neighborhoods (rows) and the enrichment of all unique cell types (columns). Neighborhoods 5,7,8, and 10 are of particular interest (Immune cells interacting with tumor cells, can be indicative of immune response to cancer).

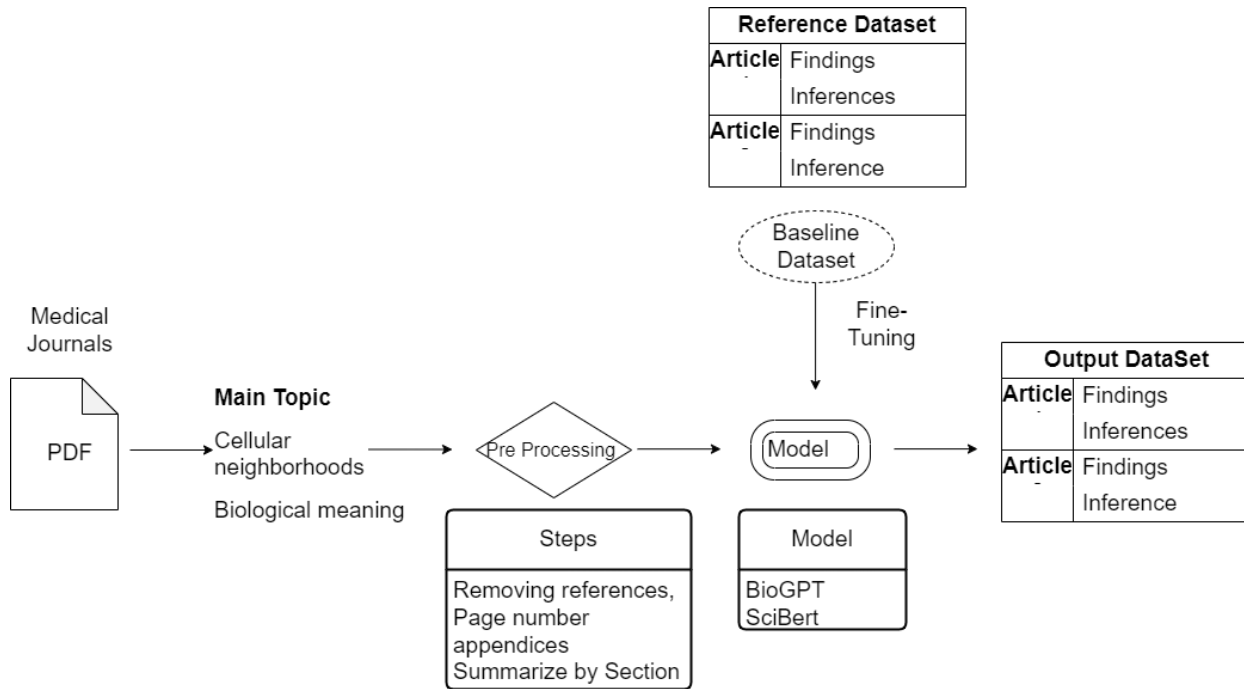


Correlations of neighborhoods with response and non-response (pre and post treatment). CN's 5 and 8 are associated with response (immune-tumor interaction), while 10 is associated with non-response (immunosuppressive cell type)

Related Work

- **Enhancing Precision in Detecting Severe-Immune Related Adverse Events [3]**
 - Here authors generate leverage a RAG-based LLM approach for taking medical records of patients taking immunotherapies which have been labeled with ICD codes.
 - They prompt models with the query record and the most similar document spans in the embedding space and evaluate accuracy based on alignment of the model prediction with existing ICD codes.
 - We can leverage a similar approach with a small labeled subset of our data to validate our modeling framework
- **BioRag [4]:**
 - Leverages RAG-enhanced LLMs to perform biological question answering using a large paper corpus and modality specific databases
 - Primary focus is question answering examples could be “given a gene, state its function”
- **Large Language Models for Scientific Information Extraction: An Empirical Study for Virology [5]:**
 - Here authors analyze article abstracts relating to COVID-19 to extract information like research problem, location, study date and reproductions number for transmittable disease
 - To validate their work, they manually annotate a gold standard subset of abstracts and fine-tune their LLM model on this subset to perform the extraction task.

High-Level Schema



Expected Outcome

Table		
Paper	Cancer type	<u>Cell Types Involved</u> Tumor Microenvironment Components <u>Spatial Arrangement</u> Cellular Interactions
Paper 1	Cancer type 1	<u>Cell Types Involved</u> Tumor Microenvironment Components <u>Spatial Arrangement</u> Cellular Interactions
Paper 2	Cancer type 2	<u>Cell Types Involved</u> Tumor Microenvironment Components <u>Spatial Arrangement</u> Cellular Interactions
Paper 3	Cancer type 3	<u>Cell Types Involved</u> Tumor Microenvironment Components <u>Spatial Arrangement</u> Cellular Interactions

Evaluation Metrics

- Our works
 - Binary classification
 - Including information about niches or not
 - Relation between two cell types (cell to cell interaction)
 - CD8+ T cell and MHC class I
 - CD4+ T cell and MHC class II
 - Multi-label Classification
 - Cell types
 - T cells
 - B cells
 - PBMCs
 - Platelets
 - Type of immune response
 - Adaptive immune response
 - Innate immune response
 - Inflammatory response
- Perplexity
 - Text summerization

References

- [1] Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat Methods* 18, 9–14 (2021). doi: <https://doi.org/10.1038/s41592-021-01065-y>
- [2] Phillips, D. et al. Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma. *Nat Commun* 12, 6726 (2021). doi: <https://doi.org/10.1038/s41467-021-26974-6>
- [3] Sun, Virginia H., et al. "Enhancing precision in detecting severe immune-related adverse events: Comparative analysis of large language models and international classification of disease codes in patient records." *Journal of Clinical Oncology* (2024): JCO-24.
- [4] Wang, C. et al. BioRAG: A RAG-LLM Framework for Biological Question Reasoning. Preprint at <https://doi.org/10.48550/arXiv.2408.01107> (2024).
- [5] Shamsabadi, M., D'Souza, J., & Auer, S. (2024). Large Language Models for Scientific Information Extraction: An Empirical Study for Virology. *arXiv preprint arXiv:2401.10040*.

2. Maanya, Jerry, Alex

Dissecting Omnibus Bills

By: Maanya, Jerry, Alex

Motivation

- Every year state governments across the country push a last-minute bill at the end of the legislative calendar. These bills are called “omnibus bills” because of the large amount of legislation they contain.
- Many omnibus bills are comprised of previously failed legislation, but it is not always clear how much of the bill this applies to.
- There needs to be some quantitative connection established between omnibus bills and previous legislation.



Similar Work

- **Tactical Maneuvering on Omnibus Bills in Congress**
 - Political Science x Logit Regression Model
 - Leader-Member Relationship
 - Advance bills that face opposition or promote party agenda
 - Bills that face opposition in Congress are 30% more likely to be attached to omnibus bills.
 - Congress-President Relationship
 - Used to avert a veto or gain support for bill
 - Bills opposed by the president are 39% more likely to be attached to an omnibus bill.
 - Environmental Factors
 - Budget deficits, divided government, issue fragmentation
 - Bills from fragmented issue areas are 37% to 42% more likely to be attached to omnibus bills

Similar Work Continued

- Learning Bill Similarity with Annotated and Augmented Corpora of Bills
 - Classifies bills in the following classes: Identical, Almost Identical, Related, Partially Related, and Unrelated.
 - Ground truth was done from human annotators as well as creating synthetic data.
 - Models used: BERT, LEGAL-BERT, RoBERTa, and SWAlign alignment algorithm.
 - Enhanced by comparing bills in the same lobbying report

Legiscan Dataset

- Legiscan is a real-time legislative tracking service.
- Has an API to pull data on every bill in state governments and Congress
- Our data containing will focus on bills in the state governments of PA, VA, MN (where we are all from)
- Data contains the text of the bill as well as metadata such as the sponsors, committee references, full history, and roll call info to help track the history of omnibus attachment bills.

Our Approach/Method

For our project we will use the legiscan dataset to achieve the following goals.

- Create a summary for each bill to group bills into different categories and aid in the task of determining the content of the legislation
- Establish similarity scores between bills to determine the progression for a bill over time. Do bills fail multiple times before they succeed? Is there a common progression that bills take over time (i.e. a few failures and then success?)
 - Here we will use ngram shingles to establish similarity scores between bills. Using these similarity scores we should be able to track the history of specific pieces of legislation over time.

Our Approach/Method part 2

- Determine the extent to which “omnibus” bills are similar to pieces of legislation proposed in the previous year(s). From this we can determine how much each omnibus bill is comprised of previously failed legislation
 - For each state in our analysis we will identify the omnibus bill and consider the similarity score between that bill and previous legislation in the same session. There are a few established metrics for determining similarity between pieces of legislation, hopefully one will work well for us. For example, Kim et al 2021 establish a 5 point scale for similarity between different bills, other papers develop similar methods.

Output/Evaluation?

Final Report

- Will include Visual representation of the progression of several bills from first proposal to passage. Something similar to a flowchart that outlines what changed in the text and when.
- For our four omnibus bills of interest, we will report the similarity scores for those bills with all other legislation proposed in the same calendar year. This should give us a good idea of exactly how much of these bills are previously failed legislation.

Evaluation

- Concrete evaluation metrics are hard in this unsupervised setting but we hope to establish good connections between bills that should be self-evident upon inspection.
- Looking for suggestions for other metrics we could use here.

3. Exploring Cultural Bias in LLMs through Connections Games

Yushui, Yifang, Zhuochun

What is Connections Game?

- Players are given a pool of words and must group them into sets of four based on a common theme or topic.
- **Example:**
 - *Pool of Words:* "Taco, Salsa, Kimchi, Chopsticks, Tamale, Ramen, Burrito, Sushi"
 - *Correct Groupings:*
 - Mexican Food: "Taco, Salsa, Tamale, Burrito"
 - Asian Food: "Kimchi, Chopsticks, Ramen, Sushi"

MALBEC	LISP	HAWKING	AETHER
EINSTEIN	RUBY	EREBUS	SYRAH
CURIE	FORTRAN	THEMIS	ZINFANDEL
HYPERION	SWIFT	RIESLING	TESLA

Project motivation

- Large language models (LLMs) can perpetuate cultural biases, reinforcing harmful stereotypes in their outputs.
- Identifying and measuring the degree of cultural bias in an LLM becomes an important factor in developing equitable models.
- This project addresses the need for evaluating cultural bias in LLMs not just through knowledge retrieval but through reasoning with culturally-based tasks.
- **Value:** Highlighting biases will lead to better, more equitable LLMs by encouraging diversity in training data and better evaluation methods.

Related work

- Samadarshi et al. (2024) and Merino et al. (2024) explored abstract reasoning in LLMs using the *Connections* game but focused on English and ignored cross-cultural evaluation.
- **Our contribution:** Extending the work by adapting the game into Chinese, analyzing how LLMs handle different cultures and languages.

Dataset

- Create a new Chinese version of the *Connections* game with 50-100 games.
- Each game contains 16 words, grouped into 4 topic categories.
- We will also provide an English translation for comparison.

Create four groups of four!

MOBILE	FOLLOWERS	SHOVELS	BUFFALO
LIKES	INSULTS	SHARES	SHEEP
APARTMENT	BILLINGS	PUPPETS	OPTIONS
EQUITY	PHOENIX	STOCKS	LEMMINGS

CONFORMISTS FOLLOWERS, LEMMINGS, PUPPETS, SHEEP
COMPANY OWNERSHIP OFFERS EQUITY, OPTIONS, SHARES, STOCKS
U.S. CITIES BILLINGS, BUFFALO, MOBILE, PHOENIX
WHAT "DIGS" MIGHT MEAN APARTMENT, INSULTS, LIKES, SHOVELS

An example of unsolved and solved connections game

Approach

- Evaluate popular LLMs (e.g., GPT-3.5 Turbo, GPT-4, LLaMA) on how well they group culturally specific words.
- Compare performance between Chinese and English versions.
- Use **Connections Game** to assess reasoning and cultural knowledge in LLMs

Model Evaluation

- Evaluation Metrics:
 - Precision and Recall: To measure the accuracy of word groupings.
 - **BERTScore**: Evaluates the similarity between the model's guessed topic and the true topic.
- Expected Outcomes:
 - We expect the results to show potential biases in LLMs, with models performing better on culturally familiar datasets and worse on unfamiliar ones.

Next Steps

- Finalize dataset creation.
- Run experiments with LLMs.
- Analyze cultural biases based on model performance
- References:
 - Tim Merino, Sam Earle, Ryan Sudhakaran, Shyam Sudhakaran, and Julian Togelius. 2024. Making new connections: LLMs as puzzle generators for the New York Times' Connections word game. arXiv preprint arXiv:2407.11240.
 - Prisha Samadarshi, Mariam Mustafa, Anushka Kulkarni, Raven Rothkopf, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Connecting the dots: Evaluating abstract reasoning capabilities of LLMs using the New York Times Connections word game. arXiv preprint arXiv:2406.11012.
 - Jialin Li, Junli Wang, Junjie Hu, and Ming Jiang. 2024. How well do LLMs identify cultural unity in diversity? arXiv preprint arXiv:2408.05102.

4. John, Rojin, Xianglong

Predicting Code-Switching Points in Chinese-English Conversations: A Comparative Study of BiLSTM + CRF and mBERT

John Bowen
Rojin Taheri
Xianglong Xu

Why Predict Code-Switching?

- Code-switching: Alternating between two languages, common in bilingual conversations
- Understanding where and why people switch languages has significant applications in multilingual NLP, including automatic speech recognition (ASR), language modeling, and dialogue systems.
- Project focus: Predict code-switching points in spontaneous Chinese-English bilingual conversations

What Have Others Discovered?

- **Blom and Gumperz (1972):** Found that low-frequency words or specialized terminology in one language often trigger switches to the other language
- **Poplack (1980):** Identified the equivalence constraint, where code-switching is more likely at points where the syntactic structures of the two languages align
- **Myers-Scotton (1993):** Code-switching often occurs at syntactic boundaries, such as between noun phrases or after prepositions
- **Fricke, Kootstra, & Meyer (2016):** Suggested that code-switching serves as a cognitive strategy, often triggered by lexical access difficulties when a word is harder to retrieve in the dominant language

How Are We Tackling This?

Data

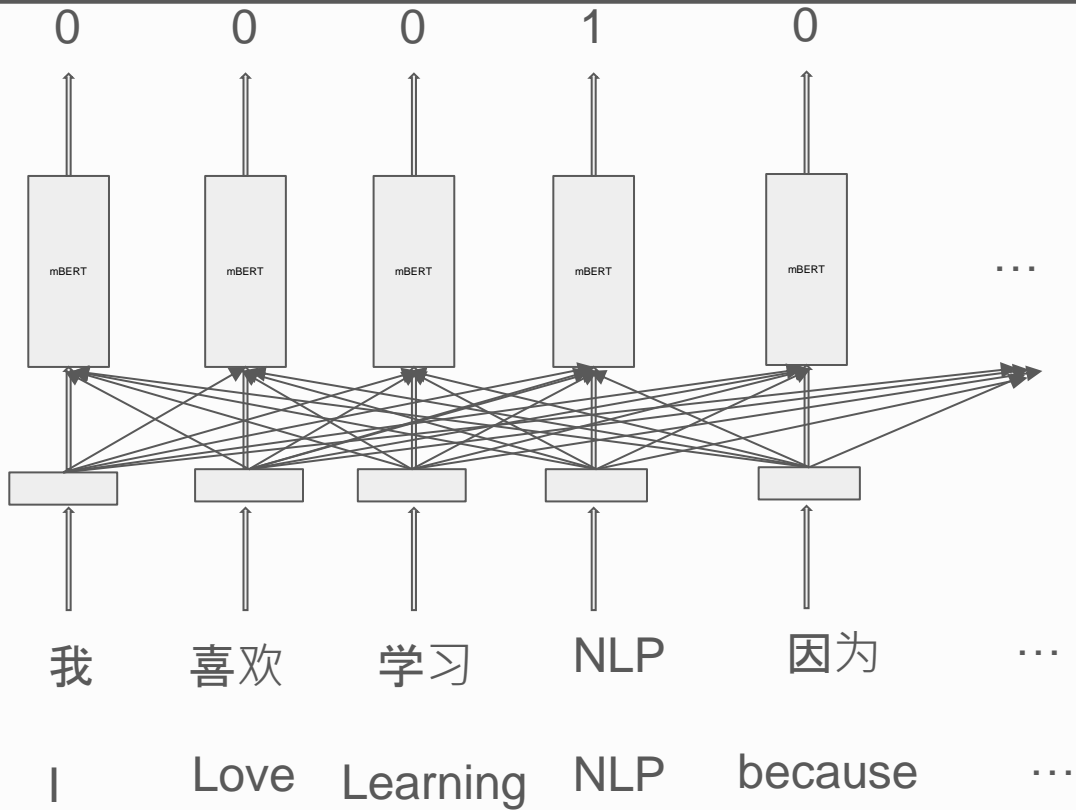
- ASCEND: A dataset of spontaneous Chinese-English code-switching conversations
- 12,314 utterances collected from 23 bilingual speakers in Hong Kong, Taiwan, and Mainland China
- Conversations cover topics like technology, education, and casual discussions
- Speaker metadata includes information on language proficiency and bilingual dominance, which will be explored as factors affecting code-switching

How Are We Tackling This?

Methods

1. BiLSTM + CRF:
 - BiLSTM (Bidirectional Long Short-Term Memory): A type of RNN that processes text in both forward and backward directions to capture context from both past and future words
 - CRF (Conditional Random Fields): A layer that ensures global consistency in labeling sequences, making it especially effective in labeling tasks like code-switching points
1. mBERT (Multilingual BERT): A transformer-based model pre-trained on over 100 languages, including Chinese and English, captures long-range text dependencies and cross-lingual patterns, making it ideal for code-switching tasks requiring context from both languages.
2. Exploration of preliminary input sentence filtering utilizing a logistical regression model to quickly identify sentences likely to contain code switching
 - a. Aimed at reducing amount of data processed by more complex models for the sake of improving computational efficiency

- The next word after switch point normally back to the original language so we can use the attention mechanism to build the relationship between each token.
- Since we don't need to predict next token, we can calculate the query for the current word with the words before and after it .



How Are We Tackling This?

Evaluation

1. Metrics:

- Precision, recall, F1 score: Evaluates accuracy in predicting code-switching points
- Token-level accuracy: Measures classification accuracy of each token

1. Model comparison:

- Analyze which model handles intra-sentential and inter-sentential switching better
- Identify model weaknesses: Missing switches related to syntax or speaker-specific traits

5. Yansheng, Xiaoyan, Shijai

https://docs.google.com/presentation/d/1VcH6a_C-wRj3-GwiH7y3fNbqoIIOZ7clwL7mNIPCeDU/edit?usp=sharing

• **Movie Summarization**

• **Based on Subtitles**

Project Motivation

- Problem Statement: Movie plot summaries are crucial for understanding and recommending films, but manually creating them is time-consuming.
- Value of Work: Automating movie summaries can save time and provide structured data for recommendation systems and review tools.
- Summary: The project aims to automatically generate movie summaries from subtitles to enhance content consumption experiences.

Related Work

- Paper 1: Liu et al. proposed a hybrid model combining extractive and abstractive methods to first extract key sentences, then generate summaries.
- Paper 2: Situmeang et al. worked on movie summarization using Indonesian subtitles with stages like preprocessing and sentence ranking.
- Our Difference: We focus on English movie subtitles for extractive summarization.

Dataset

- Data Used: CMU Movie Summary Corpus and Opensubtitles subtitle data.
- Dataset Size: Contains 1,322 movie subtitles with an average of 1,255 dialogues per movie.
- Data Processing: Metadata removal (timestamps, subtitle numbers) and cleaning of stop words and punctuation.

Methods

- Model: BERTSUM (an extractive summarization model based on BERT)
- Extractive Summary: The model selects the top-k important sentences to create a short version of the movie.
- Logic Connectors: A list was created to help prioritize summarizing, causal, and comparative sentences.

Evaluation

- Quantitative Evaluation: ROUGE-N and ROUGE-L metrics were used to evaluate the summaries.
 - ROUGE-1: 45.2%, ROUGE-2: 30.5%, ROUGE-L: 40.8%
- Qualitative Evaluation: Five reviewers rated the summaries for coherence, relevance, and quality with an average score of 3.8/5.
- Summary: The model performed reasonably well but needs improvement in capturing detailed narratives.

Conclusion & Contributions

- Dataset Contribution: We collected a dataset of 1,322 movie subtitles, which can be expanded in the future.
- Next Steps: Implementing the abstractive summarization part to enhance summary fluency and completeness.

6. Dilip, Anveshika, Akshat

Comparative Analysis for de-identification of Protected Health Information in Electronics Health Records

Anveshika Kamble

Dilip Teja

Akshat

Purpose and Motivation

Comparative Analysis for de-identification of Protected Health Information in Electronics Health Records

Protected health information (PHI) is any information in medical record that can be used to identify an individual that was created, used in course of health care service.

- Preserving confidentiality and privacy
- EHR data provides opportunities for efficient machine learning applications
- The task of de-identifying PHI is time consuming and existing softwares are limited
- Current systems in use are unable to provide clinically Inclined context based output.
- Data Utility and Availability

PHI identifiers

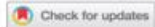
18 HIPAA PHI categories (45 CFR 164.514)

<p>1 Names;</p> <p>2 All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly-available data from the Bureau of the Census:</p> <p>a) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and</p> <p>b) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.</p> <p>3 All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;</p>	<p>4 Telephone numbers;</p> <p>5 Fax numbers;</p> <p>6 Electronic mail addresses;</p> <p>7 Social security numbers;</p> <p>8 Medical record numbers;</p> <p>9 Health plan beneficiary numbers;</p> <p>10 Account numbers;</p> <p>11 Certificate/license numbers;</p> <p>12 Vehicle identifiers and serial numbers, including license plate numbers;</p> <p>13 Device identifiers and serial numbers;</p> <p>14 Web Universal Resource Locators (URLs);</p> <p>15 Internet Protocol (IP) address numbers;</p> <p>16 Biometric identifiers, including finger and voice prints;</p> <p>17 Full face photographic images and any comparable images; and</p> <p>18 Any other unique identifying number, characteristic, or code.</p>
---	---

Related Work

- Report scores for each HIPAA PHIs and subcategories within them
- F-measure
- Re-identification
- Replacing information
- Substitutes and surrogate generation techniques

APPLIED ARTIFICIAL INTELLIGENCE
2020, VOL. 34, NO. 3, 251–269
<https://doi.org/10.1080/08839514.2020.1718343>



A review of Automatic end-to-end De-Identification: Is High Accuracy the Only Metric?

Vithya Yogarajan, Bernhard Pfahringer, and Michael Mayo

Department of Computer Science, The University of Waikato, Hamilton, New Zealand

ABSTRACT

De-identification of electronic health records (EHR) is a vital step toward advancing health informatics research and maximizing the use of available data. It is a two-step process where step one is the identification of protected health information (PHI), and step two is replacing such PHI with surrogates. Despite the recent advances in automatic de-identification of EHR, significant obstacles remain if the abundant health data available are to be used to the full potential. Accuracy in de-identification could be considered a necessary, but not sufficient condition for the use of EHR without individual patient consent. We present here a comprehensive review of the progress to date, both the impressive successes in achieving high accuracy and the significant risks and challenges that remain. To best of our knowledge, this is the first paper to present a complete picture of end-to-end automatic de-identification. We review 18 recently published automatic de-identification systems -designed to de-identify EHR in the form of free text- to show the advancements made in improving the overall accuracy of the system, and in identifying individual PHI. We argue that despite the improvements in accuracy there remain challenges in surrogate generation and replacements of identified PHIs, and the risks posed to patient protection and privacy.

Related Work

Journal of the American Medical Informatics Association, 24(3), 2017, 596–606
doi: 10.1093/jamia/ocw156
Advance Access Publication Date: 31 December 2016
Research and Applications



OXFORD

Research and Applications

De-identification of patient notes with recurrent neural networks

Franck Deroncourt,^{1,*} Ji Young Lee,^{1,*} Ozlem Uzuner,² and Peter Szolovits¹

¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA,

²Computer Science Department, University at Albany, SUNY, Albany, NY, USA

Corresponding Author: Franck Deroncourt, 32 Vassar St, 32-293, Cambridge, MA 02139, USA, E-mail: francky@mit.edu;

Tel: +1-443-637-2659

*These authors contributed equally to this work.

Received 25 June 2016; Revised 6 September 2016; Accepted 6 October 2016

Abstract

Objective: Patient notes in electronic health records (EHRs) may contain critical information for medical investigations. However, the vast majority of medical investigators can only access de-identified notes, in order to protect the confidentiality of patients. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) defines 18 types of protected health information that needs to be removed to de-identify patient notes. Manual de-identification is impractical given the size of electronic health record databases, the limited number of researchers with access to non-de-identified notes, and the frequent mistakes of human annotators. A reliable automated de-identification system would consequently be of high value.

Materials and Methods: We introduce the first de-identification system based on artificial neural networks (ANNs), which requires no handcrafted features or rules, unlike existing systems. We compare the performance of the system with state-of-the-art systems on two datasets: the i2b2 2014 de-identification challenge dataset, which is the largest publicly available de-identification dataset, and the MIMIC de-identification dataset, which we assembled and is twice as large as the i2b2 2014 dataset.

Results: Our ANN model outperforms the state-of-the-art systems. It yields an F1-score of 97.85 on the i2b2 2014 dataset, with a recall of 97.38 and a precision of 98.32, and an F1-score of 99.23 on the MIMIC de-identification dataset, with a recall of 99.25 and a precision of 99.21.

Conclusion: Our findings support the use of ANNs for de-identification of patient notes, as they show better performance than previously published systems while requiring no manual feature engineering.

Key words: medical language processing, de-identification, neural networks

- Used Bi-directional LSTMs
- Comparative analysis between CRF, ANN and LSTMs
- Word2vec and Glove embeddings

Dataset

i2b2 De-identification Dataset (Informatics for Integrating Biology and the Bedside :

Data from the i2b2 challenge, which includes 1304 clinical records of diabetic patients. These records were manually annotated by trained professionals for PHI categories such as names, dates, locations, and contact information. (Labelled)

RAW MEDICAL NOTE

Physician Discharge Summary Admit date: 10/12/1982 Discharge date: 10/22/1982 Patient Information Jack Reacher, 54 y.o. male (DOB = 1/21/1928). Home Address: 123 Park Drive, San Diego, CA, 03245. Home Phone: 202-555-0199 (home). Hospital Care Team Service: Orthopedics Inpatient Attending: Roger C Kelly, MD Attending phys phone: (634)743-5135 Discharge Unit: HCS843 Primary Care Physician: Hassan V Kim, MD 512-832-5025.

MIMIC-IV(latest version)

Structured EHR dataset grouped into two modules: hosp, and icu

Plausible Approach

- Model incorporation for The PHIs recognition :
- LLAMA (fine-tuned for medical NER) - Base model
 - Multi-class model
 - BERT (pre-trained and fine-tuned for PHI detection)
 - Anonymization(For Semantic Transformation):
MIXTRAL

Plausible Techniques (Qualitative analysis):

- Masking :

LABELLED DE-IDENTIFIED MEDICAL NOTE

Physician Discharge Summary Admit date: 10/12/1982 DATE Discharge date: 10/22/1982 DATE Patient Information Jack Reacher PATIENT , 54 AGE y.o. male (DOB = 1/21/1928 DATE). Home Address: 123 Park Drive, San Diego, CA, 03245 LOCATION . Home Phone: 202-555-0199 PHONE (home). Hospital Care Team Service: Orthopedics Inpatient Attending: Roger C Kelly STAFF , MD Attending phys phone: (634)743-5135 PHONE Discharge Unit: HCS843 HOSPITAL Primary Care Physician: Hassan V Kim STAFF , MD 512-832-5025 PHONE .

- Redacting :

DE-IDENTIFIED MEDICAL NOTE

Physician Discharge Summary Admit date: Discharge date: Patient Information , y.o. male (DOB =). Home Address: . Home Phone: (home). Hospital Care Team Service: Orthopedics Inpatient Attending: , MD Attending phys phone: Discharge Unit: Primary Care Physician: , MD .

Evaluation Metrics

- Cross-Dataset Evaluation
- Error Analysis
- Generating scores for each HIPAA PHIs and subcategories within them
- Analysis of Core Evaluation Metrics Across each Mode: (Precision, Recall, F1 Scores and Accuracy)
- Qualitative analysis (MIXTRAL)

7. Joel, Jiyang

Project Motivation

- Prominent LLM availability enables potential symbolic music generation via text-based music notation
- A new (2024) LLM benchmark for evaluation of musical understanding shows that GPT-4 outperforms music PhDs
 - Including at music generation with ABC notation
- ABC notation is old, newly relevant, and has room for exploration with modern NLP techniques

Related Work – GPT-2 Article

- 2020 - Interacting with GPT-2 to Generate Controlled and Believable Musical Sequences in ABC Notation
 - Trained GPT-2 on 300,000+ existing ABC notation single-voice tunes

Related Work - ZIQI-Eval Article

- 2024 - The Music Maestro or The Musically Challenged, A Massive Music Evaluation Benchmark for Large Language Models
 - 14,000 multiple choice questions
 - Tested 165 people; PhDs > masters > undergrad; Music majors > music background > no music background
 - **Base model GPT-4 beat every group including on ABC generation**
 - "Results indicate that all LLMs perform poorly on the ZIQI-Eval benchmark, suggesting significant room for improvement in their musical capabilities"

ZIQI-Eval Results

Educational Qualifications and Major	Music Major?	With Music Background?	Score Ranges	Average Score (Precision)
High-school	✗	✗	21 - 51	34.50
Undergraduate	✗	✗	26 - 62	39.91
Master	✗	✗	29 - 60	45.80
Ph.D.	✗	✗	26 - 51	44.00
Undergraduate	✗	✓	27 - 70	45.83
Master	✗	✓	51 - 65	57.75
Ph.D.	✗	✓	64	64.00
High-school	✓	✓	26 - 71	37.81
Undergraduate	✓	✓	30 - 88	51.52
Master	✓	✓	35 - 89	64.75
Ph.D.	✓	✓	28 - 88	64.91
GPT-4	-	-	-	67.54

Table 3: Comparison between GPT-4 and humans with different musical backgrounds and educational qualifications.

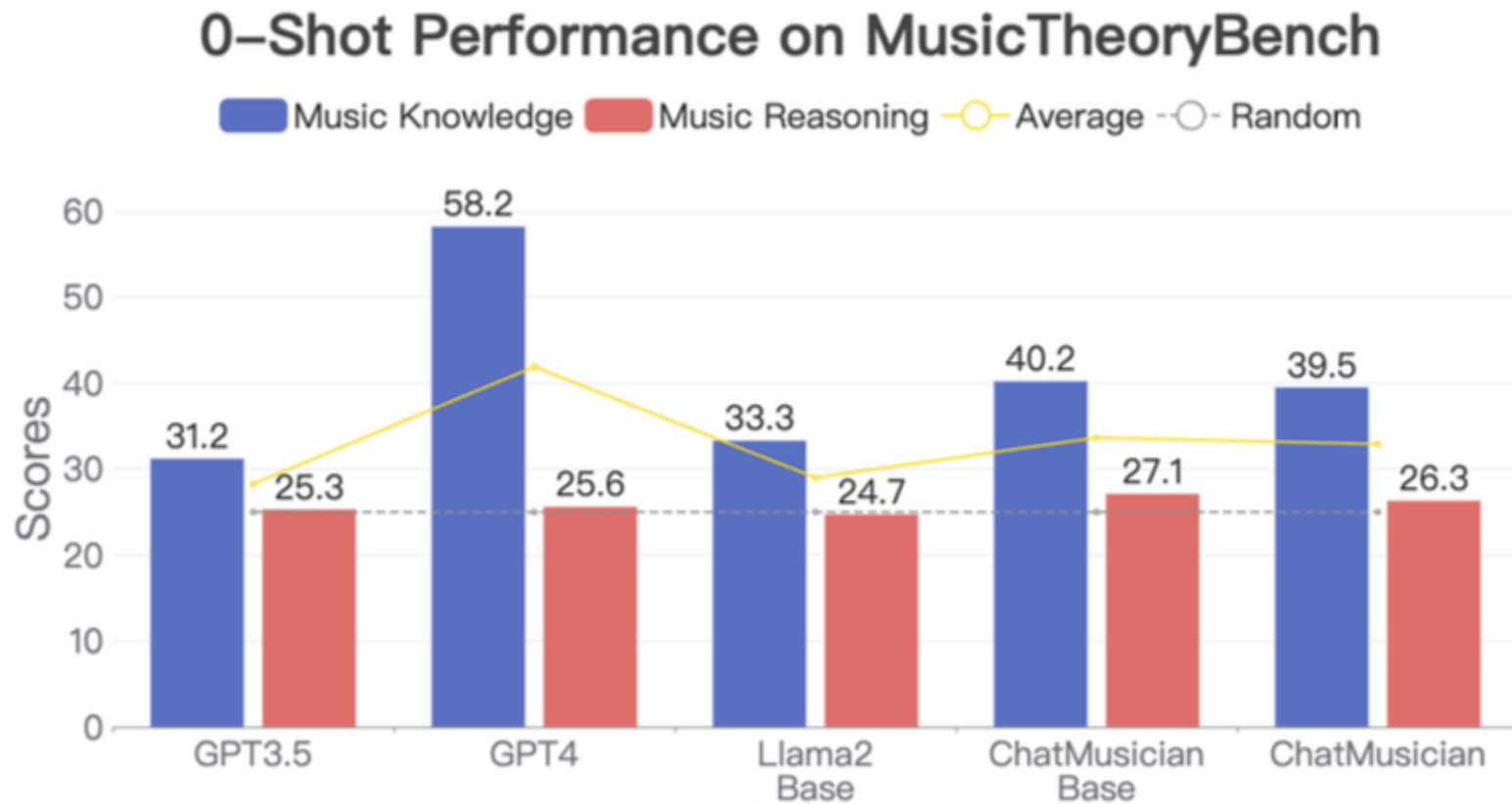
	master	Ph.D.	GPT-4
Western Music History	66.28	64.20	75.00
Popular Music	54.78	50.00	81.82
World Ethnic Music	48.54	51.52	61.11
Chinese Traditional Music	61.62	68.94	50.00
Chinese Music History	81.38	79.72	71.43
Female Music	53.29	52.27	75.00
Black African Music	82.89	72.73	100.00
Musical Performance	78.29	68.18	100.00
Music Education	52.63	63.64	50.00
Music Aesthetics	42.54	51.52	50.00
Composition Theory	63.16	67.53	50.00
Film Music	46.05	27.27	100.00
Music Generation	7.89	18.18	25.00

Table 4: Comparison of performance between highly educated individuals and GPT-4 in each category.

Related Work - ChatMusician Article

- ChatMusician: Understanding and Generating Music Intrinsically with LLM
 - Trained Llama 2 to create a custom ChatMusician model
 - ChatMusician Base was pre-trained with MusicPile
 - ChatMusician was additionally fine-tuned with supervised learning
 - MusicTheoryBench – 372 multiple-choice questions
 - GPT-3.5 and GPT-4 did about as well as ChatMusician on music reasoning

MusicTheoryBench Model Performance





Data - Sources

- 29,000 ABC notation tunes from abcnotation.com
 - Maybe 10k after dedup and cleaning
- 427 LLM generated tunes from OrchestralAI
 - Maybe more with cleaning
- 300,000 from the 2020 GPT-2 article?
 - Major cleaning likely needed

Data – ABC notation

Enter ABC Notation

```
abc
X:1
T:Ocean Raiders Overture
C:Orchestral
M:4/4
L:1/8
Q:1/4=160
K:Dm
% The introduction evokes the call to adventure on the high seas
V:1 clef=treble
%%MIDI program 71
|:"Dm" A4 A2fe | "C" d6 c2 | "Bb" B4 B2AG | "A7" A6 z2 |
"Dm" f4 fefa | "Gm" g4 gbag | "A7" a4 a2gf | "Dm" d8 :|
V:2 clef=treble
%%MIDI program 70
|:"Dm" d4 d2c2 | "C" A4- A2A2 | "Bb" G4 G2F2 | "A7" E4 E2c2 |
"Dm" A2d2 | "Gm" d2B2 | "A7" G2E2 E4 | "Dm" f4 "A7" A2G2 | "Dm" d8 :|
V:3 clef=bass
%%MIDI program 32
|:"Dm" D,4 D,2F,2 | "C" C,6 G,2 | "Bb" B,,4 B,,2D,2 | "A7" A,4- A,2z2 |
"Dm" D,8 |
"Dm" D,4 D,2D,2 | "Gm" G,,4 G,2G,2 | "A7" A,,4 A,2E,2 | "Dm" D,,8 :|
```

Paste  Edit 

Clear Upload a file

◀ ▶ 🔊 0:00

Ocean Raiders Overture

Orchestral

Musical score for Ocean Raiders Overture, tempo 160. The score is in 4/4 time and D minor. It features three staves: Violin 1 (treble clef), Violin 2 (treble clef), and Bass (bass clef). The score includes notes, rests, and dynamic markings. Colored boxes highlight specific elements: a blue box around the 4/4 time signature, a yellow box around the first measure of the Violin 2 staff, a purple box around a sixteenth-note triplet in the Violin 1 staff, and a green box around the repeat sign at the end of the Bass staff. Arrows from the ABC notation interface point to these elements.

Approach

- ABC notation data sources will need cleaned for training
- Data will be divided into training and testing data
- We will try fine-tuning multiple LLMs (ChatMusician, LLaMa3, GPT, etc.) and assessing perplexity on the hold-out set
- Expected conversation:

```
{  
  "instruction": "Construct melodies by blending the designated musical pattern with the supplied motif.",  
  "input": "['Binary', 'Sectional: Verse/Chorus'];X:1 L:1/16 M:2/4 K:G ['G2BG A2cA B2dB', '(gf)(ge) (ed)(cB)' </s> ",  
  "output": "Assistant: X:1 L:1/16 M:2/4 K:G G2BG A2cA | B2dB G2B2 | c2ec B2dB | ABAG (GF)(ED) | G2BG A2cA | B2dB c2ec | cBAG D2f2  
| g2d2B2G2 || (gf)(ge) (ed)(cB) | (gf)(ge) (ed)(cB) | ca2c Bg2B | ABAG GFED | G2BG A2cA | cBAG d2f2 | g2d2B2G2 || </s> "  
}
```


• Questions?
