FA25 Foundations of Data Curation, UIUC MCS Online

CS598 Course Project Final Report: End-to-End Data Curation Workflow

Team: Jeptha Davenport (wjd4@illinois.edu)

Project Summary Scientific examination of human consciousness has increased publication volume across fields and remains to be organized.[1] The project undertaken for this course aims to create a database of relevant publications in the field of consciousness derived from PubMed, an online database of over 39 x 10^6 biomedical literature citations maintained by the National Library of Medicine, National Center for Biotechnology Information[2] (NLM/NCBI) division. This project is limited in scope to using the first four steps of the USGS lifecycle model[3] for a cumulative dataset (Plan, Acquire, Process, and Analyze). It develops a partial mixed (manual and programmatic) data curation workflow for a 10^4 to 10^6 count citation dataset, which as a whole would be at the high limit of manual-only review, for example for purposes of meta analysis or knowledge synthesis; this could be applied to other biomedical issues of interest; this particular use case demonstrates features of data curation concerns such as data quality, automated screening, and provenance.

Artifacts generated are accessible publicly at the following repository: https://github.com/wjdavenport/FDC-Project-Backup

One-time dataset storage occurs in a pinned submodule, for archival purposes: https://github/wjdavenport/FDC-Project-Data

Metadata and Documentation: Minimal descriptive metadata for the dataset is provided in a json file[4], which follows the core elements of the DataCite 4.4 schema[5] (identifier, title, creator, publisher, resource type, subjects, dates, rights, and related identifiers). Please refer to the Project Proposal and Project Progress Report documents previously submitted and archived in the repository for additional preceding details.[6] Figure 1 gives a broad outline of the project.

[1] See Kuhn RL. A landscape of consciousness: Toward a taxonomy of explanations and implications. Prog Biophys Mol Biol. 2024 Aug;190:28-169. doi: 10.1016/j.pbiomolbio.2023.12.003. Epub 2024 Jan 26. PMID: 38281544

[2] https://pubmed.ncbi.nlm.nih.gov/, retrieved 2025-11-29

[3] J.L. Faundeen, T.E. Burley, J.A. Carlino, D.L. Govoni, H.S. Henkel, S.L. Holl, et al., "The United States Geological Survey Science Data Lifecycle Model," 2013, as cited in: Plale, Beth & Kouper, Inna. (2017). The Centrality of Data: Data Lifecycle and Data Pipelines. 10.1016/B978-0-12-809715-1.00004-3., p. 95

[4] https://github.com/wjdavenport/FDC-Project-Backup/blob/main/datacite_pubmed_consciousness.json

[5] https://schema.datacite.org/meta/kernel-4.4/, retrieved 2025-12-04

[6] Project Proposal: https://github.com/wjdavenport/FDC-Project-Backup/blob/main/Proposal%20and%20Guide/FA25%20FDC%20Proposal%20wjd4%20v03.pdf
Project Progress Report: https://github.com/wjdavenport/FDC-Project-Backup/blob/main/Proposal%20and%20Guide/FA25%20FDC%20Project%20Progress%20Report%20wjd4%20v02.pdf

Figure 1. Outline of Project Flow

PLAN

Define scope and inclusion criteria for "consciousness"-related literature

Estimate publication volume; outline workflow stages

- Consider ethical, legal, and policy constraints:
    - Copyright, licensing, OA vs restricted materials
    - Responsible use of PubMed and downloaded metadata
- Identify relevant data models and abstractions:
    - Lifecycle model (USGS: Plan → Acquire → Process → Analyze)
    - Record-centric, metadata-first modeling (schema.org/DataCite alignment)
- Anticipate metadata and documentation needs:
    - Provenance logging, dataset-level metadata, codebook
- Address reproducibility/transparency from the outset:
    - Scripts, directory structure, version control (Git/GitHub)
    - Clear documentation of assumptions and decisions

ACQUIRE

Programmatic PubMed search (1843–2025)

Download selected subset with MEDLINE metadata

- Apply FAIR principles:
    - Findable: stable identifiers, DOIs
    - Accessible: documented retrieval procedures
    - Interoperable: standard field mapping, UTF-8, JSON/CSV/SQLite
    - Reusable: licensing notes, inclusion/exclusion criteria
- Capture provenance for acquisition:
    - Search query strings, timestamps, API versions, counts returned
- Maintain documentation for ethical reuse of external data sources
- Compare to initial R script to similar e-search (~65,000 articles)

PROCESS

[Stage 1: Human Labeling + Classifier Bootstrapping]

Manually review & label 300 records

Train Classifier v1

Compare predictions vs. human labels

Identify mismatches: 38/300

Human re-review of 38 → 5 corrected

Retrain → Classifier v2

[Stage 2: Generalization Test]

Classifier v2 labels 200 additional records

Human overrides 16/200

[Stage 3: Gold Standard Consolidation]

Merge 300 (corrected) + 200 (corrected) = 500 human-verified samples

Train Final Classifier (v3)

- Data modeling & abstraction considerations:
    - Label schema, controlled vocabularies, classification conventions
    - Use of SQL/CSV structures, normalized metadata fields
- Reproducibility notes:
    - All changes logged (mismatches, corrections, overrides)
    - Versioned scripts for labeling, training, evaluation

ANALYZE

Evaluate Final Classifier:

- AUC
- Confusion matrix
- Top positive/negative n-grams

- Transparency & Reproducibility:
    - Archive scripts, parameters, random seeds, environment specs
    - Provide workflow documentation linking each result to inputs
- Metadata and documentation:
    - Dataset description, labeling criteria, variable definitions
    - Provenance trail (Plan → Acquire → Process → Analyze)
- Dissemination & communication:
    - GitHub repository or packaged ZIP archive
    - Narrative project report with methods, limitations, next steps

<u>Plan</u> The scope of this project focused data available from the publicly accessible PubMed database based on feedback at the stage of the Project Progress Report. Future project extensions could incorporate other databases, internet-based searches of online-only materials, or unpublished materials, but the current project is limited to the PubMed source. In the **Acquire** stage the date range in years was explored flexibly as detailed in the Progress Report, then fixed for a PubMed search from 1843 to 2025 (retrieval date fixed at 2025-09-27). Scripts using R[7] in RStudio[8] were created to complete acquisition programmatically. A later, parallel add-on acquisition was run as a check to the custom programmatic acquisition, based on ideas introduced in feedback[9] to the interim Progress Report for this project. **Process** involved converting the downloaded selected citations from .medline format to .csv format for subsequent manual labeling, review, and refining the analysis. This was accomplished programmatically using Python[10]. From this point, processing and **Analysis** were revisited in cycles of 'error correction' to refine selection of citations as described in those subsections of this report. This draft of the final report is submitted for review by the course instructors for comments on correctness and completeness. During coding for each of these steps, environment specifications were recorded within the repository along with code; code revision milestones were captured through versioning.

<u>Acquire</u> A programmatic acquisition was accomplished following the National Library of Medicine Entrez Programming Utilities (E-utilities) guidelines.[11] Limits followed included not making more than three URL requests per second and performing the requests during off-hours (weekends and 21:00-05:00 Eastern Time); this is noted as a policy constraint. Although the search was performed with an active ORCID (available through no-cost registration[12]) through an NCBI api key (also available at no-cost with registration[13]), it would be possible for subsequent researchers to perform the same search without registration. To avoid overly large download blocks (i.e., respecting the 10k idlist limit), a helper routine was created with an output log of the date-count ranges used to keep within such a limit. The R code used is kept at: acquire/r-project/01_download_pubmed_consciousness.R within the public GitHub repository[14]

[7] Version 4.4.1, R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.)

[8] Version 2024.09.1+394 (2024.09.1+394), Posit team (2024). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

[9] https://github.com/wjdavenport/FDC-Project-Backup/tree/main/Feedback

[10] Python Software Foundation. (2024). Python (Version 3.12.2) [Computer software]. https://www.python.org/

[11] https://www.ncbi.nlm.nih.gov/books/NBK25497/#chapter2.Usage_Guidelines, retrieved 2025-10-26

[12] https://orcid.org, retrieved 2025-10-26

[13] https://support.nlm.nih.gov/kbArticle/?pn=KA-05317, retrieved 2025-10-26

[14] https://github.com/wjdavenport/FDC-Project-Backup/

as well as on a personal storage device[15]. Using this method as described, the total number of publications captured in the PubMed search was just over 65,000, within the Project Proposal estimates (< 10^6), and these publications were the object of subsequent processing and analysis.

Pubmed records include descriptive metadata (title, abstract, authors), administrative metadata (PMID, journal abbreviation identifiers, publication dates), and structural metadata (MEDLINE[16] field record structure). This project derived Boolean indicators and text snippets for manually-assigned relevance labels; these derived metadata enabled indexing, processing, and classifier training.

Planning was modified based on Progress Report feedback, and a separate but parallel test of acquisition through software (esearch)[17] available through NLM/NCBI[18] was executed over the same date range (1843 to 2025-09-27)[19]. This acquisition showed 65,814 results, showing at least a rough quantitative equivalence in the process using R and the process using esearch.[20] Additional feedback suggested considering the PubMed query language and MeSH headings, so esearch count queries were made for that subset of 'consciousness' citations in the date range which had MeSH coverage (yielding 54,536), and there were 7,465 (or about 13.7%) which had 'consciousness' as a MeSH Major Topic.[21] This percentage was roughly similar to the original analysis using a custom classifier (vide infra, **Analysis** section).

The R download script (01_dowload_pubmed_consciousness.R) has been updated so that its default output directory matches the raw data layout in the 'data' submodule (data/raw/2025-09-27_pubmed_consciousness), ensuring that the recorded paths reflect the actual location of the MEDLINE file used in the project.

Process A Python script was created to convert the .medline file to .csv format (https://github.com/wjdavenport/FDC-Project-Backup/consciousness-ezr/scripts/01_export_pubmed_csv.py). Table 1 shows a summary of the schema used, and a data dictionary

[15] MacBook Air, M2, 2022, with 24GB memory and >1TB storage available at the time of the acquisition, processing and analysis

[16] https://www.nlm.nih.gov/medline/medline_home.html, retrieved 2025-11-30.

[17] Installation initiated via terminal as ```sh -c "$(curl -fsSL https://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/install-edirect.sh)”```

[18] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2025 Nov 29]. Available from: https://www.ncbi.nlm.nih.gov/

[19] ```esearch -db pubmed -query "consciousness AND 1843:2025/09/27[dp]" | efetch - format xml > entrez_results_01.xml```

[20] Archived at https://github.com/wjdavenport/FDC-Project-Data/blob/main/raw/entrez_results_01.xml

[21] esearch -db pubmed -query "consciousness AND 1843:2025/09/27[dp] AND medline[sb]" | grep “<Count>”; esearch -db pubmed -query “Consciousness[MeSH Major Topic] AND 1843:2025/09/27[dp] AND medline[sb] | grep “<Count>”

is stored in the data
repository.[22]

Table 1. Summary of metadata schema utilized in Process (.medline to .csv)

Field	Type(s)	Description
PMID	Identifier	PubMed-specific record ID
Year	Integer	Year of publication (range 1843 to 2025)
Journal	Text	Journal title (or source) from PubMed record
Title	Text	Journal article title
Abstract	Text	Optional, author-produced journal article summary
Medical Subject Headings (MeSH)	Formally pre-defined labels, text	Created by the National Library of Medicine, https://www.nlm.nih.gov/mesh/meshhome.html
Publication Type	Text	PubMed file publication type label, e.g., 'Journal Article; Review'
Derived feature: 'n_mesh'	Integer	Count of MeSH terms extracted for the article

As anticipated manual abstract analysis showed that a subset of flagged and acquired references was not relevant to the field of consciousness as opposed to the use of a MeSH as an indicator of level of arousal of human or animal study subjects. For example, many flagged studies referred to pharmacological, physiological or other interventions upon "conscious [rats, mice, sheep, monkeys, subjects]" while others referred to the presence or absence of consciousness as an indicator of coma or as a sequela to traumatic brain injury.

Based on the presence of such articles which ideally would be excluded from subsequent analysis, a system of parsing was developed beginning with a random sample of 300 flagged articles (pulled programmatically using https://github.com/wjdavenport/FDC-Project-Backup/consciousness-ezr/scripts/02_make_seed_labels.py) which were the manually reviewed and annotated as 'irrelevant=0' or 'relevant=1' in an expanded .csv file (https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/data/exports/seed_label_batch_working_copy_06.csv). This file was the first used to train a classifier, and from this point onward, analysis of output was re-incorporated into a further re-labeling and re-training cycle.

Subsection: Detailed discussion of the Process of human labelling (Begin)

At this point in this report it may be worth providing some examples of what constitutes 'relevant' labeling for a PubMed citation with reference to the occurrence of the word 'consciousness', and what constitutes 'irrelevant' labeling. First, it is I, the human author of this

[22] https://github.com/wjdavenport/FDC-Project-Data/blob/main/raw/2025-09-27_pubmed_consciousness/data_dict_for_pubmed_consciousness

report, who is making this determination on a subjective basis, which introduces many known and likely unknown biases into this process, but by providing a rationale as well as example heuristics as an approximation to an ideal algorithm, I hope that my method will be transparent enough that other human authors (or, in attempts such as in this project, computer models) could achieve some degree of replication using the process as well as agree in principle to my intent if not my method, and that the method may be modified as seen fit.

As alluded to above (in noting that many citations are identified in which their authors are relaying results of experiments on human subjects or animals who or which are ‘conscious’ as opposed to ‘unconscious’, e.g., anesthetized or sedated in some way, or comatose or minimally conscious—the adjectives are numerous), the presence or absence of ‘consciousness’ in such a setting would not considered ‘relevant’ to the question of *why consciousness arises at all*. I suggest that where unconsciousness is spoken of as an alternative to consciousness or a conscious state, we may assume that the content of the reference already assumes ‘consciousness’ as a pre-existing, pre-defined phenomenon and does not refer to the root problem of why certain blobs of matter seem to exhibit consciousness and others do not. This is a crude framework, incomplete and imperfect, but a scaffold that may be illuminated with specific examples I have gleaned during the course of manual reviews of model classification.

To start with an example of a ‘true positive’ in which I believe the contents of a citation are relevant to the discussion of why consciousness arises at all, and in which a model assigns a ‘more likely than not’ value to such relevance as I am training it toward (this should not be taken as evidence of correctness but is provided only to demonstrate label agreement between human and model), we have PMID 23509248, titled “Cortical response tracking the conscious experience of threshold duration visual stimuli indicates visual perception is all or none,” and for which we have an abstract to consider:

> “At perceptual threshold, some stimuli are available for conscious access whereas others are not. Such threshold inputs are useful tools for investigating the events that separate conscious awareness from unconscious stimulus processing. Here, viewing unmasked, threshold-duration images was combined with recording magnetoencephalography to quantify differences among perceptual states, ranging from no awareness to ambiguity to robust perception. A four-choice scale was used to assess awareness: "didn't see" (no awareness), "couldn't identify" (awareness without identification), "unsure" (awareness with low certainty identification), and "sure" (awareness with high certainty identification). Stimulus-evoked neuromagnetic signals were grouped according to behavioral response choices. Three main cortical responses were elicited. The earliest response, peaking at approximately 100 ms after stimulus presentation, showed no significant correlation with stimulus perception. A late response (approximately 290 ms) showed moderate correlation with stimulus awareness but could not adequately differentiate conscious access from its absence. By contrast, an intermediate response peaking at approximately 240 ms was observed only for trials in which stimuli were consciously detected. That this signal was similar for all conditions in which awareness was reported is consistent with the hypothesis that conscious visual access is relatively sharply demarcated.”[23]

[23] Sekar K, Findley WM, Poeppel D, Llinás RR. Cortical response tracking the conscious experience of threshold duration visual stimuli indicates visual perception is all or none. Proc Natl Acad Sci U S A. 2013 Apr 2;110(14):5642-7. doi: 10.1073/pnas.1302229110. Epub 2013 Mar 18. PMID: 23509248; PMCID: PMC3619304.

This example already complicates the crude framework above because the authors make reference to a hypothetical single subject (rather than an entity that is conscious and another that is not) who is aware of some things and not aware of other things. Without recapitulating schools of thought regarding consciousness here, I will say that 'conscious awareness' as distinguished 'from unconscious stimulus processing' captures a relevant issue for consciousness (of the sort I am interested in here), insofar as it seems (intuitively, introspectively, and indeed empirically) that in what is very likely a substrate for consciousness (a human brain), some processes are associated with (the experience of) consciousness, and others are not. This publication is touching on themes relevant to consciousness, I posit.

The next example turns out to be the most common sort in labeling in this project, a 'true negative,' in which I and the model both label a citation as 'irrelevant' to the discussion of consciousness I am interested in choosing (again, that the model 'agrees' with my labeling is not germane to the correctness of the label, only to an an analysis of how often the model labels things as I would have liked it to). I use example PMID 23634339, titled "Ethical Dilemma and Management of Infertility in HIV Seropositive Discordant Couples: A Case Study in Nigeria," with abstract:

> "The traditional African society places an invaluable premium on procreation and, in some communities, a woman's place in her matrimony is only confirmed on positive reproductive outcome. Infertility is rife in Nigeria, and HIV/acquired immunodeficiency syndrome (AIDS) infection is a global pandemic, which has led to a drop in life expectancy across the world. In Nigeria, a number of cultural norms relating to gender roles and power dynamics constitute a serious barrier to issues of sexuality and infertility. Couples are concerned about their infertility diagnostic test being disclosed to each other, especially before marriage. This concern is understandable, especially in an environment that lacks the modern concepts and attitude toward sexual matters. This is complicated by the advent of HIV/AIDS infection and the societal mind-set that look at seropostive individuals as transgressors. At present, sexual and reproductive health rights are currently not in place because ethical issues are not given prominence by many physicians in Nigeria. A case of an infertile and seropostive discordant couple, which raised a lot of medical and ethical concerns, is presented here to awaken the ***consciousness*** [emphasis added] of Nigerian physicians and stimulate discussions on the ethical matters such as this in clinical practice."[24]

This example highlights a use of 'consciousness' [bolded and italicized in the abstract above] which is likely much more frequent in common parlance but not in the sense I am interested in, a use implying the focus of attention of an entity already assumed to be endowed with 'consciousness'. It would be tempting to say that 'consciousness' as seen in this abstract is synonymous with 'awareness', but we have already seen that awareness versus unawareness can also be relevant to consciousness of the sort I'm trying to capture. I suggest that this sort of reference to 'consciousness' is not relevant to the question of why consciousness exists at all.

Now I will turn to an example of a 'false positive' use of 'consciousness,' in which I do not believe the selected citation bears on the question of consciousness as I would like to define it. This constitutes a 'mismatch', in which I (human) and model, do not label identically. This

[24] Umeora O, Chukwuneke F. Ethical Dilemma and Management of Infertility in HIV Seropositive Discordant Couples: A Case Study in Nigeria. Ann Med Health Sci Res. 2013 Jan;3(1):99-101. doi: 10.4103/2141-9248.109460. PMID: 23634339; PMCID: PMC3634234.

example is PMID 1319639, titled "[Diagnosis, clinical course and prognosis of parenchymatous-ventricular hemorrhages][25]," with the abstract:

> "As many as 61 patients with hemorrhagic brain stroke and blood penetration in the ventricular system were subjected to a clinical analysis. Based on the data available, attempts were made to predict an outcome of brain stroke. The authors describe the results of studying different aspects of the diagnostic algorithm of parenchymatous ventricular hemorrhages. In accordance with the clinical and computer-aided tomography data, consciousness disturbances, occlusion hydrocephalus, secondary stem syndrome as well as the localization, volume of hematomas and the degree of the blood filling of the ventricular system may serve as diagnostic predictors of brain hemorrhages. The authors' observations correspond with the conclusions made by foreign scientists that ventricular hemorrhages are not always fatal. Parenchymatous ventricular hemorrhages are likely to eventuate in a favourable outcome owing to the drug treatment."[26]

This is an example which I feel does not contribute to a discussion of the existence, occurrence or explanation of consciousness but which the model (subject to re-training) flagged as 'relevant'. Here the authors are concerned with the outcome of hemorrhagic strokes, and one of the pertinent factors for this is whether the affected person has a disturbance of consciousness. My classification here stems from the idea that the fluctuations in levels of arousal, comatose state, or alertness, to use other somewhat similar terms, presume consciousness at some prior time, and so are not 'relevant' as I categorize information about consciousness here.

The next two examples are of 'false negative' labeling by a model, in which I categorize citations as 'relevant' to discussion of consciousness of this sort which the model did not flag as relevant. I elected to provide two example here rather than one as the second may be a mismatch which has a more apparent remedy than the first. Also, the first four examples were the first four occurrences of their types (true positive, true negative, false positive and false negative) encountered in a random selection of 200 citations in a testing set (that is, not from the seed set). I point this out to demonstrate an effort to avoid bias in choosing examples, (apart from this second example in this category and the final example to follow). For PMID 31535889, titled "Evaluative conditioning of pattern-masked nonwords requires perceptual awareness," with abstract:

> "The evaluative conditioning (EC) phenomenon is central to the study of preference acquisition and attitude formation. Early studies have reported EC in the absence of awareness, but more recent work has questioned this conclusion. In previous work, using briefly presented and pattern-masked conditioned stimuli (CSs), we found that above-chance forced-choice identification of CSs is necessary for EC. Here we extend this work by ***addressing more directly the inherently subjective issue of consciousness*** [emphasis added]. In 2 studies, we assessed whether above-threshold perceptual awareness of CSs is necessary for EC. Contrasting unconscious learning claims, EC was absent under low and intermediate levels of perceptual awareness. Additional findings suggest that the perceptual awareness task does not interfere with

[25] The paper's title has been translated into English, an indicator which, we will see later, affects model prediction accuracy.

[26] Iakhno NN, Arkhipov SL, Mironov NV, Shmyrev VI, Gavrilov ES. Diagnostika, techenie i prognoz parenkhimatozno-ventrikuliarnykh krovoizliianiĭ [Diagnosis, clinical course and prognosis of parenchymatous- ventricular hemorrhages]. Zh Nevropatol Psikhiatr Im S S Korsakova. 1992;92(1):17-21. Russian. PMID: 1319639.

> EC, and that it is more sensitive than memory-based awareness proxies. We also found that a confounded variant of the forced-choice identification task can artifactually induce EC; and that an unconfounded version of the task does not induce nor interfere with EC. We discuss limitations of the present studies as well as their relevance for the debate about the automaticity of evaluative learning. (PsycInfo Database Record (c) 2020 APA, all rights reserved).”[27]

My reasoning here is similar to that in our earlier ‘true positive’ example, in that attempts at understanding what generates the ‘awareness’ in ‘consciousness’ in distinction to those processes which do not generate ‘awareness’ are relevant to understanding why consciousness occurs at all; also, the authors explicitly refer to the concept (I claim), as bolded and italicized in the abstract above.

The chosen example of a second ‘false negative’ is for PMID 39644840, titled “How can virtual reality help to understand consciousness? A thematic analysis of students' experiences in a novel virtual environment,” with abstract:

> “Research on consciousness typically presents stimuli and records the responses that follow, to infer the intervening processes. Yet, VR allows ecological validity by giving the user freedom to continuously control their sensory input across three spatial dimensions via head and eye movement. We designed a virtual world in which the angle of view relates to the information complexity of the sensory input. We assessed its acceptability and feasibility, and explored the first-person experience. Ten university students were immersed in two different novel environments, then a semi-structured interview, guided by first-person video footage of the VR experience, elicited participants' reports. The methodology proved feasible, and a thematic analysis was consistent with Mansell's (2024) control theory perspective, and to a lesser degree, Integrated Information Theory (IIT) and Global Workspace Theory (GWT). We conclude that novel virtual environments provide an accessible, dynamic and valid way to gather evidence regarding different theories of consciousness.”[28]

Despite references within the abstract by name to two current published hypotheses on the origin of ‘consciousness’ (IIT and GWT), the model did not flag this citation as ‘relevant,’ while I did for this reason as well as the authors’ concluding mention of ‘theories of consciousness,’ and I think that this example shows that classifier models could be given more direct instructions to factor in such references for increased accuracy.

I will provide one more example, chosen by me to illustrate that my choice of the ‘relevant’ label may differ from at least one connotation of ‘relevant’ as the term is used commonly, the connotation of usefulness. This is PMID 27829974, titled “Vendantic view on life and consciousness: BN Shanta is correct,” which has for abstract the following:

[27] Stahl C, Bading KC. Evaluative conditioning of pattern-masked nonwords requires perceptual awareness. J Exp Psychol Learn Mem Cogn. 2020 May;46(5):822-850. doi: 10.1037/xlm0000757. Epub 2019 Sep 19. PMID: 31535889.

[28] Gorman KR, Wrightson-Hester A, Landman M, Mansell W. How can virtual reality help to understand consciousness? A thematic analysis of students' experiences in a novel virtual environment. Conscious Cogn. 2025 Jan;127:103792. doi: 10.1016/j.concog.2024.103792. Epub 2024 Dec 6. PMID: 39644840.

> “The explanation for Vedanta offered by Bhakti Niskama Santa (BNS)(1) is valid from both scientific and philosophical grounds. It seems that the published critique of Gustavo Caetano-Anolles (GCA)(2) to Shanta's paper is purely emotional and does not have any valid scientific or philosophical justification. In his rebuttal to Caetano-Anolles's critique, Shanta(3) highlighted how the concept of 'Organic Whole' in Vedanta is completely different than that of Creationist Movement and Intelligent Design. Thus Caetano-Anolle's attempt to equate Vedanta with Creationist Movement and Intelligent Design is merely superfluous. This article highlights the validity of the argument made by Bhakti Niskama Shanta(1) and thus also intends to clarify why the Caetano-Anolles critique is groundless.”[29]

A few things to mention: the model and I both labeled this citation as ‘relevant’; it is in the format of a letter to the editor or comment, or of a reply to a comment; it appears more rhetorical in tone than many other sampled abstracts; it offers little evidence to support its conclusion. Nonetheless, it does seem to center on a discussion about the sort of consciousness I am interested in capturing, as, following this publication to the source of its commentary, I arrive at a paper[30] more obviously tied to my concept of interest. In any case, although the exemplar itself may seem at face value to have little to add to a larger discussion, it is included (by me, at least) as ‘relevant’. (Repetitively, that the model also labeled this as ‘relevant’ should not weigh in outside assessment of the correctness of this labeling.)

Subsection: Detailed discussion of the Process of human labelling (End)

Analyze Moving on from the labeling examples to some of the many challenges raised in classification attempts, I discuss the first pass in analysis. Seed labels were applied in a “bag-of-words”[31] linear text classifier implemented in Python (https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/scripts/03_train_baseline_and_ezr.py) using text feature of titles and abstracts as available (TfidfVectorizer, a scikit-learn module)[32] and modelled via binary logistic regression (LogisticRegression, also a scikit-learn module)[33]. The choice of this type of classification scheme was to preserve interpretability (since positive- and negative-

[29] Jagannadham MV. Vendantic view on life and consciousness: BN Shanta is correct. Commun Integr Biol. 2016 May 18;9(5):e1183855. doi: 10.1080/19420889.2016.1183855. PMID: 27829974; PMCID: PMC5100653. The abstract begins: “In the past, philosophers, scientists, and even the general opinion, had no problem in accepting the existence of consciousness in the same way as the existence of the physical world. After the advent of Newtonian mechanics, science embraced a complete materialistic conception about reality. ”

[30] Shanta BN. Life and consciousness - The Vedāntic view. Commun Integr Biol. 2015 Oct 9;8(5):e1085138. doi: 10.1080/19420889.2015.1085138. PMID: 27066168; PMCID: PMC4802748.

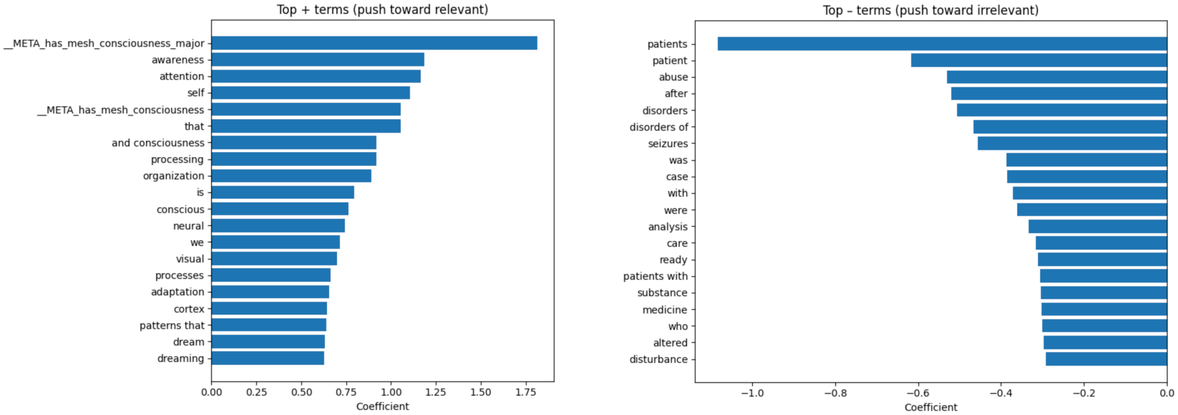
[31] https://en.wikipedia.org/wiki/Bag-of-words_model, retrieved 2025-10-26

[32] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html, retrieved 2025-10-26

[33] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, retrieved 2025-10-26

weighted n-grams are produced and preserved, as in Figure 2[34]). Figure 2 illustrates this advantage of a sparse linear classifier which was applied to the processed .csv file. Note that the n-grams varied as the classifier is trained on different training sets, in the case of this project, larger sets, as more articles were reviewed and manually labeled, providing larger numbers of possible training and testing articles. The explicit positive or negative weights contribute toward the model's probability of relevance assignment. Here we see that highly positive weights such as 'awareness', 'attention', 'visual', and 'subjective' are strongly associated with theoretical and scientific discussions of consciousness (of the sort I aim to select for). On the other hand, the strongly negative weights for 'patients', 'disorders', 'rehabilitation' correspond to clinical or biomedical scenarios that we would deem 'irrelevant' for the project. This pattern shows the classifier appears to be learning meaningful distinctions in line with the project's aim, selecting conceptual, cognitive, and philosophical terminology while eschewing other uses such as in common parlance or routine clinical communications. For additional interpretability, an additional mini-model using trees[35] was added (not shown, but appears as output to the latest classifier, as pointed to in footnote to Figure 2). Table 2 is simply a different format of the same information and lists these top positive and top negative n-grams in table form along with their co-efficent.

Figure 2. Analysis of weighted n-grams for relevance of flagged articles including "consciousness"



[34] Notes in Figure 2: this was produced by jobAnalysis.ipynb script (https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/data/labels_and_reviews/jobAnalysis.ipynb) using source files lr_tfidf_meta.joblib and tfidf.joblib (found at https://github.com/wjdavenport/FDC-Project-Backup/tree/main/consciousness-ezr/models) produced by script https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/scripts/03_train_baseline.py; the figure shown reflects the latest classifier used after training and testing on 500 manually labeled articles.

[35] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier, retrieved 2025-10-26

Table 2. List of top positive (‘relevant’ scope of term ‘consciousness’) and top negative n-grams. Produced in the output of https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/scripts/03_train_baseline.py.

```
Top positive n-grams (evidence for class 1):
__META_has_mesh_consciousness_major        1.8156
awareness                                  1.1850
attention                                  1.1670
self                                       1.1040
__META_has_mesh_consciousness              1.0556
that                                       1.0529
and consciousness                          0.9180
processing                                 0.9164
organization                               0.8908
is                                         0.7961
conscious                                  0.7634
neural                                     0.7428
we                                         0.7170
visual                                     0.6984
processes                                  0.6630
adaptation                                 0.6562
cortex                                     0.6450
patterns that                              0.6412
dream                                      0.6327
dreaming                                   0.6259

Top negative n-grams (evidence for class 0):
patients                                  -1.0836
patient                                   -0.6162
abuse                                     -0.5297
after                                     -0.5204
disorders                                 -0.5068
disorders of                              -0.4656
seizures                                  -0.4558
was                                       -0.3871
case                                      -0.3855
with                                      -0.3712
were                                      -0.3615
analysis                                  -0.3341
care                                      -0.3169
ready                                     -0.3108
patients with                             -0.3065
substance                                 -0.3041
medicine                                  -0.3018
who                                       -0.3004
altered                                   -0.2966
disturbance                               -0.2916
```

A second round of manual review of label mismatch (i.e., ‘false positives’ and ‘false negatives’ among the seed labels when their labels were stripped and exposed as test cases to the classifier. The script for the classifier was updated 1) to enlarge the set of interpreted boolean text features to include named theories or models of consciousness as well as commonly cited authors associated with such theories[36]; 2) to select cases of mismatch (https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/scripts/03_train_baseline_and_ezr_03.py)

[36] including the explicit inclusion of known proposed named models of consciousness (e.g., Global Workspace Theory, Mind-Brain Identity Theory, Orchestrated Objective Reduction, Panpsychism, Monism, Dualism, Electromagnetic Theory, Neural Correlates, and so on) and authors (e.g., Edelman, Crick and Koch, Dennett, Baars, McFadden, Searle, Chalmers)

which occurred when any of the 300 seed labels initially assigned did not agree with the repeated model probability assignment.

Mismatch among the seed labels in 38 instances (38/300 or about 12.7%)[37], and these occurrences were manually reviewed for possible seed-labeling error. Five citations on review were identified where the original seed label was deemed incorrect (four newly flagged 'irrelevant', one newly flagged 'relevant')[38] and these citations' labels were altered to reflect the new review and the 300 citations annotated and corrected this way were used as seed labels (https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/data/exports/seed_label_batch_working_copy_07.csv) for another script update with the goal of repeated training and creating an audit after this training (https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/scripts/03_train_baseline_and_ezr.py)

The audit consisted of pseudo-randomly (with seed) selecting 200 citations[39] not present in the seed set of labeled records programmatically labeled (column 'pred') by the updated model[40]; the same 200 citations were then manually labeled[41] with an extending column ('Human label') according to the same relevance criteria discussed above, with results shown in Table 3 (created manually from data obtained in the manual labeling process).

[37] https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/data/exports/seed_model_audit_01_mismatch.csv

[38] Citation ID and explanation for re-labeling as follows: Labels changed to 'irrelevant' from 'relevant': PMID 19726003 (disorders of consciousness as related to psychiatric illness, not theory of consciousness); PMID 229559 (voluntary intoxications with tricyclic antidepressants); PMID 18175088 (G-protein-coupled receptors and anesthetic effects); PMID 22928842 ("distinguishing disorders of consciousness from disorders of communication", i.e., comatose states versus locked-in states); label changed to 'relevant' from 'irrelevant': PMID 19059583 ("conscious judgment of movement" … "based on efferent … and re-afferent components", i.e., timing of conscious awareness in relation to measurable brain activity).

[39] https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/data/labels_and_reviews/manual_review_sample.csv

[40] https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/scripts/03_train_baseline_and_ezr.py

[41] https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/data/labels_and_reviews/unlabeled_model_sample_01_reviewed.csv

Table 3. Manually generated confusion matrix from interim labeling step

	Human label = 1	Human label = 0	**Total**
Model label = 1	12	11	23
Model label = 0	5	172	177
Total	17	183	200

Notes on Table 3 (calculations performed manually from hand labeling process):
Accuracy = 0.92 [(12 + 172)/200],
Precision = 0.522 [12/(12+11)],
Sensitivity (Recall or True positive rate) = 0.706 [12/(12+5)],
Specificity (True negative rate) = 0.941 [172/(172+11)],
Negative predictive value = 0.972 [172/(172+5)],
Balanced accuracy = 0.824 [(Sensitivity + Specificity) / 2]

The model predicted 23 of the 200 records as relevant and 177 as irrelevant. The classifier agreed with human (my) assessment in 184/200 cases (92%). There were 11 'false positives' and 5 'false negatives'. False positives included publications without abstracts, originally non-English works translated into English, or those making reference to generic uses of the word 'consciousness' unrelated to the target domain of this project. False negatives included publications that referenced major recognized theories of consciousness (e.g., Integrated Information Theory, Global Workspace Theory) or were without abstracts or otherwise missing theoretical terms not captures in the TF-IDF space. Despite these error modes, the classifier showed good specificity (94.1%) and negative predictive value (97.2%), suggesting that the model is reliable at rejecting irrelevant material. The model sensitivity of 70.6% likely reflects a narrow definition of theoretical relevance used in manual labeling.

Next the 200-record labeled set was merged with the 300-record labeled set programmatically[42] to create a 'gold-standard' file[43] for the latest training and testing. The classifier[44] was run to show current performance, storing results in a small output file[45], shown in Table 4.

[42] Using script https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/scripts/04_merge_labels.py

[43] https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/data/labels_and_reviews/full_labeled_set.csv

[44] https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/scripts/03_train_baseline.py

[45] https://github.com/wjdavenport/FDC-Project-Backup/blob/main/consciousness-ezr/data/baseline_lr_metrics.csv

Table 4. Classifier performance after expanding the training and testing set.

Baseline Logistic Regression Classifier Performance
(500 human–labeled articles; 75/25 train–test split)

Metric	Value
Training samples	375
Test samples	125
Test positive rate	0.20
Accuracy	0.832
ROC AUC	0.908
PR–AUC	0.777
True positive rate (recall, class 1)	0.760
Specificity (true negative rate)	0.850
Balanced accuracy	0.805

The classifier uses a TF–IDF representation of title and abstract text, augmented with four MeSH-based meta-features, and is trained via L2-regularized logistic regression with class weighting. A stratified 75/25 train–test split yielded 375 training and 125 testing samples, with a positive-class prevalence of 0.20 in the test set. On this held-out test set, the model achieved an accuracy of 0.832, a ROC AUC of 0.908, and a PR-AUC of 0.777. The confusion matrix (TN = 85, FP = 15, FN = 6, TP = 19) corresponds to a sensitivity of 0.760 for in-scope articles, a specificity of 0.850 for out-of-scope articles, and a balanced accuracy of 0.805. Given the underlying positive-class prevalence of 0.204 (which represents the naïve PR-AUC baseline), the observed PR-AUC of 0.777 indicates that the model is substantially more effective than random or keyword-only screening. All metrics are written automatically to `baseline_lr_metrics.csv` by the script `03_train_baseline.py`, supporting reproducible evaluation.

Comments on Provenance & Reproducibility: Aligning with concepts covered in this course of provenance metadata, source identification, transformation and environment documentation, the exact query parameters, dates, files (with checksums when particularly large, e.g., hundreds of MB) were captured, version-controlled, and stored along with the environment details and scripts and software package versions in two GitHub repositories.[46] Another facet of reproducibility susceptible to variance stems from the 'human-in-the-loop' curation through seed labeling which I generated from a subjective operational definition of 'relevance' to 'consciousness'; this was compounded by my choice pursue this project as a solitary endeavour rather than with a group, where co-authorship would allow for consensus building between or among two or more project authors.

Comments on Constraints:

Copyright and reproducibility: Manual classification has been possible without recourse to bulk downloading of copyrighted references, and where questions of relevance arose (less often than 3% in sampling), single-reference review was available through institutional access (either University of Illinois Urbana-Champaign or University of Calgary) for the <100 references pulled for full review) for research and educational purposes of this project.

[46] https://github.com/wjdavenport/FDC-Project-Backup/tree/main and https://github.com/wjdavenport/FDC-Project-Data/tree/main

Policy: As noted in the **Acquire** section, bulk download restrictions for the public database (PubMed) were identified and adhered to.

Copyright and legal: The metadata available through PubMed does not include publicly available resources including books, recorded lectures, interviews, conference talks, preprints, dissertations and theses, and liberal arts fields (humanities and philosophy) of relevance. The Project focus remained on refining the current PubMed search results and analysis so that a similar selection process can be applied to such search extensions beyond this initial project. Ethics are an important concern for data curation projects, and this project makes use of abstracts provided by the National Library of Medicine, but NLM does not hold copyright on the abstracts but has been given permission to use the abstracts by their respective publishers. Some of the abstracts may be subject non-U.S. copyright restrictions. I have added a note on copyright to the two repository README documents reflecting information from the NLM[47].

F**uture work**: Expansion of this project could implement a books pilot using one or more curated bibliographies cross-matched to GoogleScholar/GoogleBooks to test a merger schema. I anticipate the number of published books relevant to the topic to be 10^2 (2 orders of magnitude less than journal references). The pilot would produce 1) merger schema (book_id, title, authors, year, publisher, ISBN/DOI, subject keywords, citations (connected to PubMed PMIDs where available); 2) provenance log (source bibliography, retrieval date, query strings); 3) sample knowledge-graph sketch (author-work-cites relationships) to demonstrate a lineage and cross-pollination of ideas in the field of consciousness studies.

Timeline: The final report was prepared after feedback on a draft received from course instructors[48], with final submission (for course grading by rubric) to occur by December 10, 2025. This report and prior versions are stored in the project repository.[49]

[47] https://www.nlm.nih.gov/databases/download.html, retrieved 2025-11-30.

[48] https://github.com/wjdavenport/FDC-Project-Backup/blob/main/Feedback/Project_draft_final_report_feedback.rtf

[49] https://github.com/wjdavenport/FDC-Project-Backup/tree/main/Proposal%20and%20Guide