Plan Overview

A Data Management Plan created using DMPonline

Title: Visible Nonsense, Hidden Sense: Extending Drivel-ology to Challenge the Pragmatic Comprehension of Large Multimodal Models

Creator: Yang Wang

Principal Investigator: Chenghua Lin

Data Manager: Chi-Li Chen, Chia-Yi Hsiao, Yiqi Liu, Zi Yan Chang

Contributor: Tyler Loakman, Chenghao Xiao

Affiliation: University of Manchester

Template: University of Manchester Generic Template

Project abstract:

While Large Multimodal Models (LMMs) demonstrate increasing proficiency in processing concurrent audio, visual, and text inputs, their ability to perform holistic, pragmatic reasoning across these modalities remains a critical and underexplored frontier. Current benchmarks often test descriptive accuracy or factual retrieval, leaving a gap in evaluating a model's grasp of layered, non-literal meaning. To address this, we introduce DrivelHubV2, a significant evolution of the Drivelology framework ("nonsense with depth"). This new benchmark is designed specifically to probe the limits of integrated pragmatic comprehension in a tri-modal context.

DrivelHubV2 is structured around a challenging task: Tri-modal Multiple-Choice Question Answering (MCQA). Each data point presents a Drivelological scenario through either image or video, accompanied by a textual question. The core challenge is not merely to describe the scene or sound, but to perform the non-linear, pragmatic, and cultural reasoning required to understand its "nonsense with depth", the same cognitive challenge posed in DrivelHubV1. The model must perceive the visual and auditory cues and then reason about their collective non-literal meaning to identify the correct underlying narrative from five options. The distractor options are deliberately crafted to represent common failure modes, such as providing a purely literal description of the visual scene or misinterpreting the audio's tone, without grasping the deeper paradoxical or subversive intent.

We hypothesise that even state-of-the-art LMMs will struggle significantly with DrivelHubV2, given that their uni-modal counterparts consistently fail to master the purely textual Drivelology in our original benchmark. The addition of audio and visual modalities serves to amplify this fundamental challenge, requiring models to move beyond surface-level perception and apply complex reasoning to a richer, multi-sensory input. By focusing on this potent task structure, DrivelHubV2 provides a precise tool for measuring whether current architectures can overcome this deep representational gap. We will release the dataset and evaluation framework to the community to catalyse research into more robust and genuinely pragmatic multimodal reasoning.

ID: 192166

Start date: 15-12-2025

End date: 15-05-2026

Last modified: 12-12-2025

Copyright information:

Visible Nonsense, Hidden Sense: Extending Drivel-ology to Challenge the Pragmatic Comprehension of Large Multimodal Models

Manchester Data Management Outline

1. Will this project be reviewed by any of the following bodies (please select all that apply)?

- Ethics

2. Is The University of Manchester collaborating with other institutions on this project?

- Yes - Part of a collaboration and owning or handling data

Other collaborators are from University of Manchester, Durham University, and University of Sheffield.

3. What data will you use in this project (please select all that apply)?

- Acquire new data

We will create **DrivelHubV2**, a new multimodal dataset. The data will consist of collected public data and created annotations. The dataset will comprise approximately 1000 publicly available short videos and images sourced from social media platforms, including YouTube, Instagram, Threads, and TikTok.

4. Where will the data be stored and backed-up during the project lifetime?

- Other storage system (please list below)

Data will be stored in a University of Manchester password protected secure OneDrive folder, which is an Information Governance Office (IGO) approved data storage system and the first choice platform for storing information.

5. If you will be using Research Data Storage, how much storage will you require?

- Not applicable

6. Are you going to be receiving data from, or sharing data with an external third party?

- Yes

Data will be shared between University of Manchester, Durham University, and University of Sheffield.

7. How long do you intend to keep your data for after the end of your project (in years)?

- 5 - 10 years

Guidance for questions 8 to 13

Highly restricted information defined in the Information security classification, ownership and secure information handling SOP is information that requires enhanced security as unauthorised disclosure could cause significant harm to individuals or to the University and its ambitions in respect of its purpose, vision and values. This could be: information that is subject to export controls; valuable intellectual property; security sensitive material or research in key industrial fields at particular risk of being targeted by foreign states. See more examples of highly restricted information.

If you are using 'Very Sensitive' information as defined by the Information Security Classification, Ownerships and Secure Information Handling SOP, please consult the Information Governance Office for guidance.

Personal information, also known as personal data, relates to identifiable living individuals. Personal data is classed as special category personal data if it includes any of the following types of information about an identifiable living individual: racial or ethnic origin; political opinions; religious or similar philosophical beliefs; trade union membership; genetic data; biometric data; health data; sexual life; sexual orientation.

Please note that in line with data protection law (the UK General Data Protection Regulation and Data Protection Act 2018), personal information should only be stored in an identifiable form for as long as is necessary for the project; it should be pseudonymised (partially de-identified) and/or anonymised (completely de—identified) as soon as practically possible. You must obtain the appropriate ethical approval in order to use identifiable personal data.

8. What type of information will you be processing (please select all that apply)?

- Audio and/or video recordings

The dataset will comprise approximately 1000 publicly available short videos and images sourced from social media platforms, including YouTube, Instagram, Threads, and TikTok. We make sure we will use API to download the data and follow T&C from each platform. We will generate textual annotations and metadata for each video and image. These annotations will describe the content of the media.

9. How do you plan to store, protect and ensure confidentiality of any highly restricted data or personal data (please select all that apply)?

- Store data in buildings, rooms or filing cabinets with controlled access
- Store data in encrypted files, folders, computers or devices
- Where needed, follow University of Manchester guidelines for disposing of personal data

10. If you are storing personal information (including contact details) will you need to keep it beyond the end of the project?

- Not applicable

11. Will the participants' information (personal and/or sensitive) be shared with or accessed by anyone outside of the University of Manchester?

- Yes - Public institutions with contractual arrangements (e.g. NHS research sites or other higher education institutions)

Dataset will be shared between University of Manchester, Durham University, University of Sheffield. There is not participant involved in this dataset, only publicly available video/image will be downloaded from social media, including TikTok, YouTube, Instagram, and Threads. We will download the data via API and follow T&C from each platform.

12. If you will be sharing personal information outside of the University of Manchester will the individual or organisation you are sharing with be outside the EEA?

- No

Dataset will be shared between University of Manchester, Durham University, University of Sheffield.

13. Are you planning to use the personal information for future purposes such as research?

- No

14. Will this project use innovative technologies to collect or process data?

- No

15. Who will act as the data custodian for this study, and so be responsible for the information involved?

Chenghua Lin will be responsible for the data management as the PI of the project. Specifically in: data capture, data quality, storage, backup, archiving and data sharing.

16. Please provide the date on which this plan was last reviewed (dd/mm/yyyy).

2025-12-12

Project details

What is the purpose of your research project?

We introduce DrivelHubV2, a significant evolution of the Drivelology framework ("nonsense with depth"). This new benchmark is designed specifically to probe the limits of integrated pragmatic comprehension in a tri-modal context.

What policies and guidelines on data management, data sharing, and data security are relevant to your research project?

Our project will adhere to the relevant policies and guidelines of the University of Manchester. Specifically, we will follow:

- The University of Manchester Research Data Management Policy
- The University of Manchester Records Management Policy
- The University of Manchester Publications Policy
- The University of Manchester IT policies and guidelines
- The University of Manchester Intellectual Property Policy
- The University of Manchester Data Protection Policy

In accordance with these policies, data will be stored on secure, backed-up servers with access limited to the research team. For sharing, the final dataset will be released to the research community under a restrictive, non-commercial license to prevent misuse.

Responsibilities and Resources

Who will be responsible for data management?

Professor Chenghua Lin, as the PI, is ultimately responsible for data management. However, the responsibilities will be divided as follows:

- Professor Chenghua Lin:
  - Will provide insight for the project management plan.
  - Will ensure all data management activities comply with university policies and legal requirements.
  - Will oversee decisions regarding long-term data preservation and sharing.
- My Role (Researcher/PhD):
  - I will be responsible for the day-to-day data management tasks.
  - This includes creating and maintaining data management plan, organising and documenting data, performing regular backups, and managin data access.

What resources will you require to deliver your plan?

The primary resources required are staff time for data curation, secure server space for storage and backup, and repository costs for long-term data sharing.

Data Collection

What data will you collect or create?

We will create **DrivelHubV2**, a new multimodal dataset. The data will consist of collected public data and created annotations.

- **Collected Data:** The dataset will comprise approximately 1000 publicly available short videos and images sourced from social media platforms, including YouTube, Instagram, Threads, and TikTok.
- **Created Data:** We will generate textual annotations and metadata for each video and image. These annotations will describe the content of the media.
- **Data Storage Volume:** With an estimated size of up to 100MB per file, the total estimated storage volume for the dataset will be approximately **100 GB**. This project will not involve any direct research participants or interview recordings.

Compliance and Ethical Considerations: In handling this data, we will strictly adhere to the following guidelines and procedures:

1. **University Policies:** All data collection and management will be conducted in accordance with the University of Manchester's **"Social Media and CCTV Guidelines for Research and Recruitment"** and the principles outlined in the **"Guidance on recordings"** on the Data Management and Protection page.
2. **Platform Terms of Service:** We will review and comply with the terms and conditions of each social media platform (YouTube, Instagram, Threads, TikTok) from which data is sourced.
3. **Anonymisation and Data Minimisation:** We are committed to protecting the privacy of individuals. During the annotation process, we will anonymise personal data (e.g., usernames, locations) found in the collected videos and images. We will only gather the necessary information required to answer our research questions, and textual descriptions will be paraphrased where appropriate to protect the identity of the original posters.

How will the data be collected or created?

Collection and Curation: Data will be sourced from the public APIs of social media platforms. We will perform a manual review to select relevant content and discard any material containing private information or hate speech.

Quality Control and Standards: To ensure high quality and consistency, we will implement several measures:

- **Annotation Manual:** All data will be annotated by a trained team following a detailed manual that provides clear definitionsy.
- **Peer Review:** A senior researcher will review a subset of the annotations for quality assurance.
- **Organisation, Documentation, and Version Control:** In line with the University of Manchester and UK Data Service guidance, we will adopt a systematic approach to data organisation:
  - **Folder Structure:** We will use a hierarchical folder structure to keep data and documentation separate and organised. For example:
    - DrivelHubV2/
      - data/ (containing subfolders for raw_videos/, raw_images/, and annotations/)
      - documentation/ (containing the annotation_manual.pdf, version_log.csv, etc.)
  - **File Naming Convention:** Files will be named consistently to ensure they are meaningful and sortable. The convention will be [DataType]_[SourceID]_[Date]_[Version].ext, using the YYYY-MM-DD format for dates. For example: video_tiktok123_2025-12-05_v1.0.mp4.
  - **Version Control:** We will use numerical versioning in filenames (e.g., v1.0, v1.1, v2.0) to track revisions. A version control table will be maintained in the documentation folder to log significant changes, dates, and authors. Final versions of files will be clearly marked as such.

Documentation and Metadata

What documentation and metadata will accompany the data?

The **DrivelHubV2** dataset will be accompanied by comprehensive documentation and metadata, created in accordance with the CESSDA Data Management Expert Guide.

1. Documentation: We will provide two levels of documentation:

- **Project-Level Documentation:** A detailed README.md file in the dataset's root directory will serve as a quick-start guide. It will describe the project's aims, data collection and processing methodologies, the directory structure, and provide code examples for linking annotations to media files. The repository will also include a LICENSE.md file and a Responsible_Use_Agreement.md that users must accept before access.
- **Data-Level Documentation:** Contextual information for each video/image (e.g., source platform, collection date) and the textual annotations will serve as the data-level documentation, providing details for each individual data item.

2. Metadata: To ensure the dataset is Findable, Accessible, Interoperable, and Reusable (FAIR), we will use a formal metadata standard.

- **Metadata Standard:** We will adopt the **Dublin Core Metadata Standard**, a widely recognised schema suitable for describing digital resources.
- **Implementation:** A metadata file (e.g., metadata.csv) will be provided, where each row corresponds to a media file and columns represent Dublin Core elements such as Title, Creator, Source, Date, Description, Format (e.g., MP4, JPG), and Rights (linking to the license). This ensures that the dataset is machine-readable and easy to integrate with other data catalogs.

Ethics and Legal Compliance

How will you manage any ethical issues?

We will manage ethical issues through a multi-layered approach, which will be submitted for full review and approval by the University of Manchester's ethics committee before any data collection begins.

Our strategy includes:

- **Public Data Only:** We will only collect content that has been made publicly available by its creators on social media platforms.
- **Content Moderation:** A rigorous manual review process will be implemented to filter out and exclude any content that contains hate speech, incites violence, or reveals sensitive personal information.
- **Anonymisation Strategy:** Following the principles of the UK Data Service Anonymisation resource and the UK Anonymisation Network Framework, we will implement a systematic anonymisation process to manage disclosure risk.
    - **When:** Anonymisation will be performed by our trained annotators as a dedicated step **during the manual data review and annotation phase**, before any data is added to the final, shareable dataset.
    - **How:** We will distinguish between direct and indirect identifiers:
        - **Direct Identifiers:** All direct personal identifiers found in the source data (e.g., social media usernames, direct links to user profiles) will be **completely removed**. They will not be stored or distributed in any form.
        - **Indirect & Visual Identifiers:** We will review the visual content of images and videos for potentially identifying information (e.g., faces of individuals, visible addresses, license plates). This information will be **blurred or obscured** unless it is essential for the research context, balancing the need for privacy with data utility.
- **Takedown Policy:** We will establish a clear and easily accessible takedown policy. If an original creator contacts us and requests the removal of their content from our dataset, we will comply immediately and without question.
- **Responsible Use Agreement:** Access to the dataset will be conditional upon users agreeing to a

"Responsible Use Agreement." This agreement will strictly prohibit any attempts to de-anonymise individuals, use the data for surveillance purposes, or engage in any form of harassment.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

Our management of IPR distinguishes between the original content and the novel annotations we create.

- **Third-Party Content:** The copyright for the original videos and images remains with their respective creators. Our use of this content is for non-commercial, academic research purposes. We are not claiming any ownership over this third-party material.
- **Generated Data (Annotations):** The intellectual property rights for the novel annotations, the dataset's structure, and the accompanying documentation will be owned by the University of Manchester, in line with institutional policy.
- **Licensing for Reuse:** To promote open and responsible research, the annotations and metadata we generate will be licensed under a **Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license**.
    - This license requires users to give appropriate credit (**Attribution**).
    - It restricts the use of the dataset to non-commercial purposes (**NonCommercial**), which respects the likely intent of the original creators.
    - It requires any derivative datasets to be shared under the same terms (**ShareAlike**).
- **Permissions:** The aforementioned Responsible Use Agreement will also require users to acknowledge the IPR of the original creators and to use the data in a manner that does not infringe upon their rights.

Storage and backup

How will the data be stored and backed up?

All project data will be stored on the University of Manchester's OneDrive for Business. This is a secure, university-approved, cloud-based storage solution.

Data is automatically backed up in multiple locations to prevent loss, and will be retained for a minimum of 10 years after the project's completion. No data will be stored on personal devices.

How will you manage access and security?

Access and security will be managed through a combination of technical controls and formal agreements, adhering to university policies.

- **Risk Management:** The primary risk is the potential misuse of the dataset. We mitigate this by controlling access, enforcing a Responsible Use Agreement, and having a clear takedown policy.
- **Access Control (During Project):** During the data collection and annotation phase, access to the raw data will be strictly limited to the named project researchers (Prof. Chenghua Lin's team). Access will be controlled via the University of Manchester's standard authentication and authorisation system, ensuring only approved individuals can view or modify the data.
- **Secure Collaboration:** All data transfers and access to the central storage will occur over encrypted connections (e.g., HTTPS, SFTP), ensuring secure access for all team members.
- **Access Control (Post-Project):** Upon completion, the curated dataset will be made accessible via the **University of Manchester's institutional data repository, Figshare** . Prospective users will be required to register and digitally sign a **Responsible Use Agreement** before being granted download access. This agreement will legally bind them to the terms of the CC BY-NC-SA 4.0 license and prohibit any attempt to de-anonymise individuals or use the data for malicious purposes.

Selection and Preservation

Which data should be retained, shared, and/or preserved?

All data that is necessary to validate our research findings and has long-term reuse value will be retained, shared, and preserved. In line with guidance from the UK Data Service, we have selected open, non-proprietary formats to ensure the data remains accessible in the medium and long term.
The data package will include:

- **The curated media files:** The final, cleaned set of video clips will be preserved as **MPEG-4 (.mp4)** files, and images as **JPEG (.jpg)** files. These are widely-used, standard formats recommended for digital video and image data by the UK Data Service, ensuring high compatibility.
- **The annotation data:** The corresponding annotations will be stored in **JSON** format. As a standard, open, text-based format, JSON is ideal for structured data, ensuring it is both machine-readable for computational analysis and human-readable for inspection.
- **Comprehensive documentation:** This includes a "Datasheet for Datasets," a **README.md** file, and a **LICENSE.md** file. Using plain text Markdown (.md) ensures the documentation is maximally accessible and can be opened on any platform without specialised software.

This entire package of data will be retained and shared to allow for the replication of our results, to support new studies in computational linguistics and social media analysis, and for potential use in teaching. In accordance with University of Manchester policy and common funder requirements, the data will be retained for a minimum of **10 years** after the project's completion. Any intermediate or raw data with no long-term value will be securely destroyed.

What is the long-term preservation plan for the dataset?

Our long-term preservation plan is to deposit the complete, curated **DrivelHubV2** dataset into the **University of Manchester's institutional data repository**.
This strategy ensures the dataset becomes a sustainable and durable resource for the research community. The plan involves:

- **Preparation for Archiving:** Before the end of the project, we will finalise the dataset and its comprehensive documentation (Datasheet, README) to ensure it is fully understandable and reusable by others.
- **Deposit in the Repository:** We will deposit the complete dataset into the university's repository, which is managed by the library and IT services. This service is designed for long-term curation and preservation, guaranteeing the data remains accessible well beyond the lifetime of the grant.
- **Persistent Identifier:** Upon deposit, the dataset will be assigned a **Digital Object Identifier (DOI)**. This creates a permanent, citable link to the data, facilitating proper attribution and making it easily discoverable by other researchers.
- **File Formats:** The chosen file formats (.mp4, .jpg, .json) are standard and stable, making them suitable for long-term preservation and future format migration if necessary.

Data Sharing

How will you share the data?

We will share the data through established, public data repositories to ensure maximum visibility, accessibility, and long-term impact for the research community. Our sharing strategy is as follows:

- **Primary Repository:** As this project is a collaboration between the **University of Manchester, the University of Sheffield, and Durham University**, the specific institutional repository for depositing the data will be determined through a **Data Sharing Agreement** led by the University of Manchester. This ensures the final, curated dataset is formally archived, receives a persistent Digital Object Identifier (DOI) for stable citation, and is preserved for the long term.
- **Community Platform:** To reach the most relevant audience, we will also host a version of the dataset on **Hugging Face**. As a central hub for NLP and machine learning datasets, this will facilitate easy integration with common model training and evaluation pipelines, significantly boosting its reuse and impact.
- **Timing:** The dataset will be made publicly available concurrently with the publication of our primary research findings, or at the end of the project grant period, whichever is sooner.
- **Audience:** The data will be shared with academic and non-commercial researchers globally.
- **Acknowledgement:** Re-use of the data will be acknowledged through the citation of the dataset's DOI and our associated research paper.

Are any restrictions on data sharing required?

Yes, specific restrictions are necessary to address Intellectual Property Rights (IPR) and ethical considerations related to the source material. The goal is not to limit research, but to ensure the data is used responsibly. The main restrictions are:

- **Non-Commercial Use:** The copyright for the original content remains with its creators, who made it public but did not consent to its commercial reuse. To respect their IPR, the dataset will be shared under a **Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY-NC-SA 4.0) license** . This legally restricts the use of the data to non-commercial purposes only.
- **Responsible Use Agreement:** Although the data is public, there is a risk of misuse. To mitigate this, users will be required to agree to a "Responsible Use Agreement" before accessing the data. This agreement will explicitly prohibit any attempts to de-anonymise individuals, use the data for surveillance, or engage in harassment.

These measures represent our strategy to share the data as openly as possible while upholding our ethical and legal obligations to the original content creators.

Planned Research Outputs

Conference paper - "Visible Nonsense, Hidden Sense: Extending Drivel-ology to Challenge the Pragmatic Comprehension of Large Multimodal Models"

Planned research output details

Title	DOI	Type	Release date	Access level	Repository(ies)	File size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Visible Nonsense, Hidden Sense: Extending Drivel-o ...		Conference paper	Unspecified	Open	None specified		None specified	None specified	No	No