Research Workflow Document

📎 **What is this for:** Consumer Motivation Analysis of “Dupe” Products (Study 1)
Author: Meng-Jen (Miya) Lin
Date: December 2025

Text-based Analysis of “Dupe” Product Consumption

Focus: Hedonic vs. Utilitarian Motivations & Purchase Intention

✏️ 1. Research Context & Objective

This workflow document describes the end-to-end analysis pipeline I implemented for my PhD research proposal on *"dupe" (alternative) product consumption*.

The primary objective of this demo is to show **how unstructured social media text** can be transformed into **hypothesis-ready data**, and specifically how I test:

> 👉 H1: Hedonic motivation is associated with higher purchase intention for dupe products compared to utilitarian motivation.

The pipeline is implemented in Python and designed to be:

- **Modular** (separable components for I/O, cleaning, and analysis)
- **Reproducible** (fixed configuration, deterministic AI labeling)
- **Extensible** (scalable to larger real-world datasets beyond the demo sample)

📁 2. Data & File Structure

2.1 Data Sources (Demo Version)

For this demonstration, I use a simulated dataset that mimics social media posts about "dupe" products (e.g., Instagram, Dcard, PTT, Facebook, TikTok).

- `data/raw/dupe_posts_sample.csv`
  - Columns: id, platform, date, text
  - Content: Short user-generated texts describing dupe-related purchases or intentions.

In the full project, this file would be replaced by a **large-scale corpus** collected via the private AI automation agent (LLM + MCP).

2.2 Processed Data

The pipeline produces two key processed files:

- `data/processed/ai_labeled_results.csv`
  - Main working dataset for hypothesis testing
  - Includes AI-generated labels for:
    - **Motivation**: Hedonic vs. Utilitarian
    - **Purchase Intention** (e.g., "bought / will buy" vs. "no intention")
- `data/processed/dupe_posts_cleaned.csv`
  - Tokenized and cleaned text, ready for topic modeling or further NLP analysis.

📜 3. Pipeline Overview

The project is organized into three main modules:

3.1 Statistical Analysis (Hypothesis Testing)

- `analysis_demo.py`

- Reads `ai_labeled_results.csv`
- Computes:
  - **Purchase rates** by motivation type
  - **Chi-square tests** for association
  - **Logistic regression** with motivation as a predictor of purchase intention

This script serves as the **minimal, transparent core** for verifying H1.

3.2 Traditional NLP Pipeline (Exploratory Text Analysis)

Located under: `src/dupe_pipeline/`

- `config.py`
  - Centralizes file paths and key parameters (e.g., input/output directories), ensuring the pipeline can be ported across environments with minimal changes.
- `data_io.py`
  - Handles robust data loading and saving in a consistent way.
- `cleaning.py`
  - Applies Regex-based cleaning and Chinese tokenization (via jieba), preparing text for frequency analysis or topic modeling.
- `run_pipeline.py`
  - Orchestrator script that:
    1. Loads raw data
    2. Runs cleaning and tokenization
    3. Produces word frequency outputs for visualization

3.3 Reporting & Visualization

- `reports/figures/word_freq_top.png`
  - Visual summary of the most frequent tokens/keywords in the corpus.
- `reports/data_for_datawrapper.csv`
  - Aggregated frequency table exported specifically for external tools (e.g., Datawrapper).

🌱 4. Methodological Evolution

(Proposal vs. Implementation)

In my original Research Proposal, I planned to use **Structural Topic Modeling (STM)** to uncover latent themes in dupe-related conversations.

In this implementation, I upgrade the methodology to a **Generative AI–based labeling approach**, for two main reasons:

1. **Semantic Precision**
   - LLMs can distinguish subtle contextual differences (e.g., *"I want to buy"* vs. *"I bought it before"*) that are often missed by classical topic models.

2. **Scalability & Reproducibility**
    - The system uses the **Groq API** serving **Llama-3.3-70b**.
    - I deliberately chose an **open-weights model** over closed models (e.g., GPT-4, Gemini) to improve transparency and long-term reproducibility in an academic context.

🧠 5. AI Labeling Mechanism (Private Automation Layer)

The **AI Automation Agent** that handles data fetching and labeling lives in a **private repository**, as it is part of an ongoing longitudinal project.

To keep this demo academically useful yet IP-safe, I expose only the *downstream* labeled data (`ai_labeled_results.csv`) and document the controlled procedure:

1. **Zero-shot Classification**
    - The LLM classifies each post into Hedonic vs. Utilitarian motivation (and purchase intention), using carefully designed prompts rather than large supervised training sets.
2. **JSON Schema Enforcement**
    - The model is required to respond in strict JSON format, making the output fully machine-readable and easy to validate.
3. **Deterministic Inference (Temperature = 0)**
    - I set the model's temperature to 0 so that repeated runs on the same input produce identical labels-crucial for scientific replication.

High-level flow (Mermaid diagram equivalent):

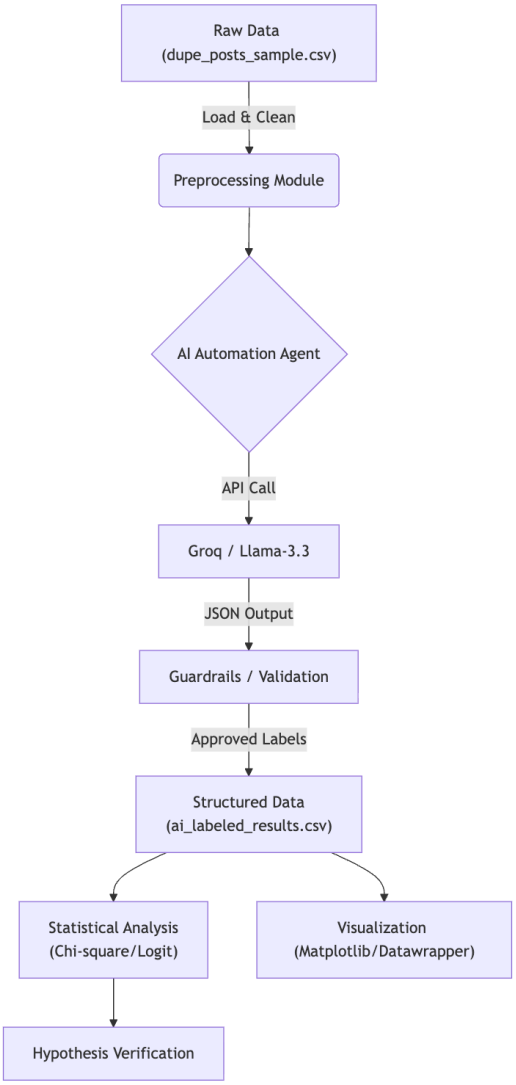

Figure 1: Computational Research Pipeline

📈 6. Statistical Analysis Summary (H1 Demo)

Using the labeled dataset, I test:

> 📎 H1: Hedonic motivation drives higher purchase intention for dupe products than utilitarian motivation.

Results (demo dataset):

Motivation Type	Purchase Rate	Sample Size
Hedonic	**91.84%**	49
Utilitarian	**50.00%**	50

- **Chi-square test:**
  - χ^2 = 20.91, p < 0.001 → Strong evidence of association

- **Logistic regression:**
  - Hedonic motivation coefficient ≈ **2.42**, indicating a large positive effect on purchase intention.

These numbers are based on the demonstration sample, but the **pipeline itself** is designed to scale to larger, real-world corpora.

🛤️ 7. Reproducibility & Environment

- Core libraries (demo):
  - numpy, pandas, scipy, statsmodels, jieba
- Environment management:
  - All dependencies are listed in requirements.txt.
- Reproducibility choices:
  - Deterministic AI labeling (temp=0)
  - Centralized paths in `config.py`
  - Clear separation between raw, processed, and report outputs.

👣 8. Planned Extensions

In the full project, I plan to:

1. **Scale up data collection** across more platforms and time windows.
2. **Refine labeling schema** to include mixed motives and contextual moderators (e.g., income, relationship status).
3. Integrate **unsupervised learning (e.g., STM)** on top of the LLM-labeled data to identify nuanced sub-themes (e.g., *"guilt-free spending"* vs. *"smart consumerism"*) within the broader hedonic segment.
4. **Extend the pipeline** to additional hypotheses (e.g., temporal framing, self/other focus) using the same automation backbone.

📲 9. Contact

For committees or collaborators interested in reviewing the **full automation agent** (including MCP integration and LLM orchestration), I am happy to provide temporary access to the private repository.

Meng-Jen (Miya) Lin

National Taiwan Normal University

Email: **mengjen.miya.lin@gmail.com**