From Archives to AI: Residential Property Data Across Three Decades in Brunei Darussalam (Data in Brief) – Response to Reviewers

*Haziq Jamil, Amirah Noorosmawie, Hafeezul Rabu, Lutfi Abdul Razak*

We thank the reviewers and editors for their thoughtful and constructive feedback on our manuscript. We appreciate the time invested in reviewing our work and have provided detailed responses below (in blue font) addressing each comment line by line.

Additionally, here is a summary of general changes to the manuscript:

* Updated data set to include January and February 2025 house listings.
* Updated LLM data extraction to latest deepseek-r1:14b reasoning model for increased accuracy.
* Re-validated data set for internal consistency and duplication removals.
* Added ‘Data Validation’ section under methods.
* Added a simple missing data pattern analysis.
* Refocus manuscript to data curation rather than economic analysis.

We hope these revisions and clarifications demonstrate our commitment to enhancing the clarity, rigour, and overall quality of the manuscript.

## Reviewer #1

The manuscript presents a valuable dataset covering Brunei's residential property market from 1993 to 2024, which is likely to serve as an important resource for researchers in economics, urban planning, and real estate. I have some thoughts on improvement that could be implemented to highlight and make clear the limitations and applications of the data created.

Firstly, the data collection approach combines manual transcription from newspaper archives, web scraping, and LLM-based cleaning. While this multifaceted method is innovative, the reliance on manual processes and Excel for spatial harmonisation raises concerns about reproducibility and scalability. Wondering if the author has thought about this - not just in terms of introducing human error - but also how it can contribute the longevity of the data source moving forward.

We thank the reviewer for raising this important point regarding reproducibility and scalability in our data curation process. We acknowledge that a fully automated pipeline would be ideal, and have given careful consideration to these issues. In our current framework, we do not anticipate further collection of "old" sources from newspaper archives, as our primary focus moving forward will be on web-scraped data. With an expected yield of approximately 1,000–2,000 entries per year (roughly 150 entries per month), the data volume is well within the capacity of our existing processes, while still offering the potential for automation. We recognise the advantages of a completely automated pipeline—such as scheduling a cron job on GitHub to routinely update the dataset—and plan to leverage increased automation in future iterations while maintaining robust quality control measures.

A particular challenge remains in the spatial harmonisation of location data. Despite our efforts to use programming for matching reported locations to our reference database, issues such as spelling variations (e.g., "Tanjong Bunut" vs. "Tanjung Bunut"), similar names with minimal differences (e.g., "Bengkurong" vs. "Sengkurong"), and ambiguous cases (e.g., "Pandan A, B, C" and "Berakas A, B") persist. These challenges are currently addressed through manual intervention by experienced checkers who leverage additional contextual clues--such as nearby road names--from the advertisement listings. Importantly, any discrepancies are largely mitigated at the aggregated mukim level.

To further enhance this process, we are exploring the development of a programmatic "black box" that employs advanced data cleaning and wrangling techniques to resolve such ambiguities with quantifiable confidence. This function could eventually be integrated into the open-source R package {bruneimap} for the benefit of the broader research community.

While a degree of manual checking remains indispensable to ensure internal consistency, our current approach incorporates systematic quality control measures that flag and correct anomalies. In sum, although our methodology presently incorporates some manual steps—particularly for spatial harmonisation—these are necessary given the non-standard nature of location reporting in Brunei. Our future plans include further automation to improve reproducibility and scalability, ensuring that the dataset remains a robust and reliable resource for analysis.

Second, since the dataset is predominantly derived from newspaper listings and online property advertisements, there is a risk that certain segments of the market—particularly properties not advertised through these channels—may be under-represented. Incorporating additional sources, such as government records (for example, land registries or records of land transactions), along with validation from real estate agents, could contribute to a more representative sampling of the residential property market.

We thank the reviewer for this thoughtful comment. We acknowledge that our dataset captures only a subset of the housing market, as it is predominantly derived from newspaper listings and online property advertisements. Selective sampling is inherent when relying on publicly available advertising data; however, this approach offers significant advantages, including lower barriers to data collection and the capacity for timely updates that facilitate effective trend analysis. Notably, many established house price indices, such as the Rightmove index in the UK, are based on similar data sources and yet reliably capture overall market trends.

While we recognise that incorporating additional sources, such as government records or land registry data—with further validation from real estate agents—would undoubtedly enhance representativeness, such integration remains a promising avenue for future research beyond the current proof-of-concept. We believe that the benefits of our approach, particularly its transparency and reproducibility, outweigh this limitation, and we are committed to exploring these enhancements in subsequent work.

Thirdly, the LLM-based cleaning process is a notable innovation in this study. However, with an accuracy rate of 93%, there is potential for errors to propagate in the final dataset. A more detailed error analysis or cross-validation with alternative cleaning methods would enhance confidence in the dataset. In addition, the operational and computational costs associated with employing LLMs should be taken into account, as these may affect the scalability of the approach.

We thank the reviewer for highlighting the innovative nature of our LLM-based cleaning process and for raising important points regarding error propagation and computational costs. At the time of the original submission, we employed the Llama 3.1 model with 8B parameters; however, since then, significant advancements in reasoning models have emerged. Following the reviewer’s suggestion, we conducted an in-depth cross-validation study using alternative models on a test dataset—details of which are now provided in the 'LLM Data Cleaning' subsection as well as the appendix.

Our experiments revealed that state-of-the-art models, such as the OpenAI o1-mini reasoning model (accessed via API), can achieve accuracy rates as high as **99.2%,** while locally run reasoning models like DeepSeek R1 (Distilled Qwen 2.5) with 14B parameters performed impressively with a **96.9%** accuracy rate.

Reasoning models are a broad category of LLMs optimised for structured thinking, logical inference, and decision-making. While reasoning models offer superior accuracy, they require longer processing times; for example, processing 5,400 entries on our high-performance computing cluster took approximately 20 hours. Cloud-based LLMs on the other hand offer even more superior speed and accuracy, but requiring a payment-based API access.

Given that our current operational needs involve processing only around 150 new entries per month—a volume that is manageable on a modern personal laptop—we are satisfied with the performance of the locally run DeepSeek R1 model.

Finally, it is important to note that the dataset relies on listing prices as a proxy for market values. Listing prices do not always reflect actual sold prices due to negotiation dynamics and marketing strategies. A factual discussion of this limitation is recommended, particularly if market trends or policy implications are to be derived from the analysis.

Thank you for the comment. We completely agree on this point, and that a key limitation of this dataset is that listing prices may not reflect final sale prices due to negotiations, seller strategies, and market conditions. Properties can sell at a discount or premium based on factors like bargaining power, financing constraints, and economic trends. While the dataset offers valuable insights into price trends, caution is needed when drawing policy implications or market forecasts. Future research could improve accuracy by incorporating actual transaction data, if available from official sources. We have now included a discussion of this limitation in the manuscript.

## Handling editor

The LLM-based data cleaning introduces subjectivity and potential bias. Some elaboration here is required.

We thank the reviewer for this comment. We acknowledge that every data cleaning process, whether conducted by human transcribers or via an LLM-based approach, inherently involves some degree of subjectivity. However, we believe that our LLM-based cleaning process offers a consistent and systematic methodology, particularly when handling thousands of listings in a **short amount of time**. In contrast, human transcription of large datasets (e.g., 10,000 listings) done **hastily** can be prone to fatigue and inconsistent interpretation, which may introduce comparable, if not greater, variability.

As suggested by Reviewer #1, we conducted a cross-validation study of the accuracy of LLMs in the data extraction process. The study, which is now described in the manuscript (LLM Data Cleaning section and appendix), demonstrated a 96.9% accuracy rate, providing further support regarding the reliability of our approach. Moreover, we have implemented robust quality control measures—namely manual verification of outliers—to mitigate potential inaccuracies introduced during the automated process. While no method is entirely free from subjectivity, the scalability and reproducibility afforded by the LLM-based approach represent significant advantages in managing large-scale data cleaning tasks.

The paper claims significance for policy applications, but the dataset is specific to Brunei, limiting generalizability. some elaboration would be appreciated.

We thank the reviewer for this comment. While our dataset is specific to Brunei, it remains highly valuable for policy applications by offering detailed insights into long-term housing trends, affordability, and spatial dynamics. Its methodological approach—combining archival and automated data collection—can be applied in other contexts. Although direct comparisons with other markets may be limited, the dataset enables cross-country studies when used alongside similar datasets, offering a useful reference for understanding housing trends in small, resource-rich economies. Moreover, Brunei’s compact geographic structure allows for comprehensive national-level spatial analysis, which is a distinctive advantage that may not be feasible in larger economies. Overall, the primary contribution of our work lies in its methodological innovation, offering a scalable model for other under-studied small economies to generate and leverage similar open datasets for policy analysis.

How was data accuracy validated (especially for manually transcribed entries)? What are the error margins for web-scraped data?

We thank the reviewer for this important comment. First, we describe the manual transcription process. This began with capturing photographs of the physical or electronic copies of the advertisement listings, which were then organised into monthly folders on a laptop computer prior to transcription. Two dedicated transcribers worked at a manageable pace—approximately 150 entries per week per person over nine months, totaling 12,210 entries—ensuring that the task was undertaken without undue time pressure. An agreed-upon data template was used to streamline the transcription process, and regular meetings were conducted to address any concerns that arose along the way.

Following this, we implemented **internal consistency checks and outlier detection,** such as conducting summary statistics analyses to identify and flag anomalous values (e.g., a built-up area recorded as 0.1 square feet instead of a realistic value, which often indicated a column mix-up). Validity checks included using the price per square foot indicator, which is known to typically fall within a specific range. As a note, these consistency checks were done applied to the human transcribed data, as well as the LLM data.

In contrast, the web-scraped data reproduces exactly the information published on the source websites, resulting in a measurement error that is effectively zero; any inaccuracies inherent in the advertisements themselves are beyond our control. The deterministic nature of the web scraping procedure—facilitated by the stable website structure during the data collection period—ensured precise extraction of the required property characteristics.

We believe that these quality control measures sufficiently address potential data accuracy issues, ensuring a robust and reliable dataset. To enhance transparency for readers and potential users of the data, we have now included a detailed description of this validation procedure in the manuscript.

How consistent is the extraction across different property descriptions? LLM outputs may suffer from data hallucination—was there a manual verification step?

Thank you for the question. Concerns regarding consistency and potential hallucinations in the LLM-based extraction process are hopefully addressed with the aforementioned cross-validation study using 100 test data points. To reiterate, our experiments demonstrated an accuracy rate of 96.9% for the LLM-based extraction, with accuracy reaching as high as 99.2% when using state-of-the-art, paid OpenAI models. These findings underscore the reliability of the extraction process despite potential hallucination issues. The experiments highlight that the greatest challenge for the LLM was accurately classifying the build status of a property—specifically differentiating among “Under Construction,” “New,” and “Resale” statuses—which appears to be attributable to the vagueness of some property descriptions. In addition, the same robust manual verification procedures described previously were applied to the LLM outputs, ensuring that any inconsistencies or errors were identified and corrected.

The LLM model and parameters are not fully described (e.g., temperature setting, prompt structure). The dataset is not fully independent of subjective processing.

Thank you for this comment. In response, we have now disclosed the LLM parameters in the appendix, including details on temperature, top‑p, top‑k, maximum tokens, and repeat penalty—most of which were set to their default values, with the exception of a lowered temperature to minimise creativity in output. Furthermore, the manuscript includes a detailed description of the LLM prompt structure (see Figure 5), which clearly outlines the specific instructions provided for data extraction, and we have also supplied a reproducible R script that users can employ to perform the data cleaning. Note that there was no system wide prompt used.

We acknowledge that any data cleaning process—whether conducted by humans or LLMs—involves some degree of subjectivity. However, the combination of systematic quality control measures, such as outlier detection and manual verification, along with the consistency achieved through our automated processing and cross-validation experiments, effectively mitigates potential subjectivity and ensures the robustness of the dataset.

and finally: focus on data curation, not economic analysis!

Thank you for the guidance. To this end we have made the following changes:

* Rewrote the ‘Value of the Data’ section. In particular, we now highlight the multi-faceted methodological innovation in data curation, and we removed the overly economic flavoured paragraph ‘Influence on economic and monetary policy’.
* Relegated the RPPI comparison to the appendix.
* Limitations section now also focus on less on economic limitations, but rather data limitations.

## Scientific Editor

1. Specifications table/data accessibility: Please make the data citable via zenodo and include a link here.

HAZIQ to do

Reviewer's Responses to Questions

1) Are these data original and produced by the authors? Please respond with Yes OR No OR N/A.

Reviewer #1: Yes

2) Are these data secondary (e.g. censuses, government databases, organizational records)? Please respond with Yes OR No OR N/A. If YES, please answer 2a, 2b & 2c; if NO go to 3

Reviewer #1: Yes - though data has been sourced from existing records, such as newspaper advertisements and online property listings, not census or records.

2a) Secondary Data Only: were these data collected using variables that make the study unique? Please respond with Yes OR No OR N/A.

Reviewer #1: Yes

2b) Secondary Data Only: is this collection of secondary data of value to the research community? Please respond with Yes OR No OR N/A.

Reviewer #1: Yes

2c) Secondary Data Only: do the authors provide the protocol for collecting/creating these data? Please respond with Yes OR No OR N/A.

Reviewer #1: Yes

3) Have the authors used a questionnaire or survey? Please respond with Yes OR No OR N/A. If YES, please answer 3a; if NO go to 4.

Reviewer #1: No

3a) Is the sampling representative of the population and rigorously following a scientific method? Please comment on the rigor of the sampling method and if additional sampling or a different sampling method is required. Please also mention if the questionnaire/survey being used is direct, unambiguous and unbiased.

Reviewer #1: The sampling method primarily relies on collecting property listings from newspaper archives and online platforms, which essentially constitutes a form of convenience sampling.

It is hard to tell if the sampling is biased to agents that actively list their properties, long-standing RE companies, and whether there are missing listings that make an impact to the overall usability and generalisability of the data produced.

If/wherever possible, it would be beneficial to incorporate additional data sources. Government records—such as those from land registries or records of land transactions—could provide an official baseline. Also, engaging with real estate agents for corroboration could help to identify any systematic biases in the listings and ensure that the data reflects actual market conditions, rather than merely the prices at which properties were initially advertised.

4) Do the authors adequately explain to the research community the utility of these data in the “Value of data” section? Please include a comment on the validity of this section. Include notes on how this can be improved, if necessary.

Reviewer #1: Yes, though could have more meat on the bones.

Thank you—this section has been rewritten based on the handling editor’s comments as well.

5) Are these data described clearly in the “Data description” section? Please provide suggestions to the author(s) on how to further clarify the presentation and description of the dataset.

Reviewer #1: Yes

6) Is the protocol/method for generating these data adequately described in “Experimental design, materials, and methods” section? Please include suggestions on how the section can be improved to aid reproducibility/reusability.

Reviewer #1: Authors should consider the following:

Authors describe several methods for data collection, the reliance on manual transcription and Excel-based spatial harmonisation raises concerns regarding reproducibility and scalability. Perhaps we need to discuss this.

Thank you—this has been addressed above.

The integration of historical data is a notable strength, yet differences between manually collected early data and later web-scraped data may introduce inconsistencies. The manuscript would benefit from further discussion on how these differences might affect comparability over time.

Thank you—this has been included in limitation section.

The use of an LLM-based cleaning method is innovative; however, with a reported accuracy of 93%, some errors may persist. A more detailed error analysis or cross-validation with alternative cleaning methods would provide additional assurance regarding the data’s accuracy.

Thank you—this has been addressed above.

Another important aspect to consider is the reliance on LLMs for data cleaning. The cost associated with deploying LLMs—both in terms of computational resources and the expertise required to maintain and update these systems—should not be underestimated. The financial and operational costs may impact the scalability of the workflow, particularly if ongoing adjustments are needed to accommodate changes in data sources or to improve model performance.

Thank you—this has been addressed above. To add to this, the open source tools such as Ollama API and associated models allow anyone with minimal skill set to access LLM data cleaning. We also provide a reproducible R script for users to perform such cleaning.

7) Have the authors provided all the raw data related to all the tables, graphs, images and charts, etc. and are they freely accessible? Please provide suggestions to the author(s) on how to improve data accessibility for wider usage. Please mention missing raw data, if any.

Reviewer #1: Yes

8) If this data article is related to an existing primary research article is there any duplication? If yes, please comment on this. Please mention any overlapping text, images, etc.

Reviewer #1: Not known