Department for Transport

AI Consultation Analysis Tool (CAT) v1.0 Evaluation

DfT AI and Data Science Team

The Alan Turing Institute

Department for Transport
Great Minster House
33 Horseferry Road
London
SW1P 4DR

This publication is also available on our website at www.gov.uk/government/organisations/department-for-transport

Any enquiries regarding this publication should be sent to us at www.gov.uk/government/organisations/department-for-transport

Contents

Table of Figures

About and acknowledgements

About the team

The CAT was co-developed between the DfT AI and Data Science team and The Alan Turing Institute.

Anton Poletaev[†] Ana Wheelock Zalaquett[‡] Chris Steinberg[‡] Harry Shepherd[‡] Ines Heimann[‡] John Francis[†] Saba Esnaashari[†]

[‡)] **AI and Data Science team, DfT Advanced Analytics (AI.Programme@dft.gov.uk)**

[†)] **The Alan Turing Institute**

Acknowledgements

Dr Jonathan Bright

Executive summary

Background

The Department for Transport (DfT) conducts around 55 consultations annually, generating large volumes of free-text responses from consultees that require thematic analysis – extracting a set of themes and systematically classifying those themes to individual responses. Completing this analysis manually is highly resource-intensive and costly. To address this challenge, the DfT and The Alan Turing Institute co-developed the Consultation Analysis Tool (CAT), an AI-powered system that completes thematic analysis with human oversight.

Development began following DfT's (2023)[1] research on public attitudes towards the use of AI for analysing consultations, which has shaped the responsible design and evaluation of the CAT. The CAT has now been piloted on multiple live consultations, analysing 200,000 responses (exceeding 8 million words) to date.

This report details the CAT v1.0 methodology, its performance evaluation results against human benchmarks, summarises key findings from its use on live consultations, and shares lessons learned. We estimate that the CAT saves around 50-70% of the entire cost and time required to respond to a medium-sized consultation (see Section 10 for details), whilst maintaining a level of quality comparable to human experts.

How does the CAT work?

The CAT uses AI to analyse free-text responses for two key stages of thematic analysis:

1. **Theme generation:** An ensemble of large language models (LLM) extracts a set of main themes and golden insights (i.e., rare themes) from free-text responses.

2. **Theme mapping:** An ensemble of LLMs classifies which human-validated themes are mentioned in each response (i.e., multi-label classification).

[1] Department for Transport. (2023). Public attitudes to the use of AI in DFT consultations and correspondence. https://www.gov.uk/government/publications/public-attitudes-to-the-use-of-ai-in-dft-consultations-and-correspondence

Following the theme generation step, a structured human theme review is conducted to validate, refine, and, where necessary, augment the AI-generated themes. This involves reading a random sample of responses and reviewing the CAT-generated themes using a structured method. This formal human oversight step is designed to validate themes and ensure no main themes are missed.

The final output is a structured dataset where every response is thematically analysed, accompanied by descriptive statistics that *estimate* theme prevalence rates by question and categories of interest (see Section 3 for details on the methodology).

How was the CAT evaluated?

Our evaluation compares CAT-analysed data to human-analysed reference data for both steps of thematic analysis – theme generation and theme mapping. We used two distinct evaluation designs which we refer to as 'blind' and 'non-blind' in this report to reflect differences in how the human-analysed reference datasets were produced and used.

- **Blind evaluation design:** This design assesses the CAT's ability to independently reproduce a human-analysed dataset, without the possibility that either the CAT or human analysts influence each other. We applied this approach to 11 questions to test the CAT's reliability and validate enhancements during development. This built confidence before evaluating CAT on live consultations.

- **Non-blind evaluation design (live pilots):** This design reflects how the CAT performed on live consultations. Human experts (policy professionals and analysts) reviewed the CAT's initial analysis and made any necessary adjustments to the thematic analysis after examining the underlying data themselves. The final human-adjusted dataset was then compared to the CAT's original output. Between 32 (theme generation) and 7 (theme mapping) questions were used at this stage.

Together, these two complementary evaluation designs demonstrate strong alignment between the CAT's analysis and human judgement, strengthening confidence in the accuracy and robustness of the CAT. The datasets are described in Section 4.

Performance evaluation findings

- **Theme generation:** Using our *blind evaluation design,* CAT-generated themes were compared to human-generated themes. The CAT detected around 75% of human-generated themes (recall) without any human oversight. Theme matches between human-generated (true) and CAT-generated (predicted) themes were determined using an evaluated LLM-as-a-judge approach. For live pilots (*non-blind evaluation design)*, initial CAT-generated themes were compared to the human-validated ones following the above-mentioned structured theme review. Overall recall was 90% in this setting. These findings evidence that the CAT performs with high accuracy in automated theme extraction, whilst human review remains important to ensure that all main themes are identified.

- **Theme mapping:** Using our *blind evaluation design*, the CAT's mapping of themes to individual free-text responses (i.e., multi-label classification) was compared to human-mapped datasets. In this setting, the themes themselves were first

generated by human researchers. The CAT achieved an overall accuracy (F_1) score of 0.75 (out of 1). For live pilots (*non-blind evaluation design*), social researchers reviewed and amended any of the CAT's theme mapping classifications that they disagreed with. The CAT achieved an overall F_1 score of 0.93 when comparing the initial CAT mapping to the human-adjusted mapping.

- **Theme mapping inter-rater reliability with human experts:** The CAT-vs-human inter-rater reliability (IRR), using metrics commonly employed in qualitative research to assess how consistently two or more researchers analyse the same data, achieved over 92% overall raw agreement in both our blind and non-blind evaluation designs. Interpreting the corresponding Cohen's kappa values, a raw agreement IRR measure corrected for random chance agreement, indicates that the CAT achieved IRR levels falling between 'substantial' and 'almost perfect' according to some standard qualitative interpretations of kappa values. In addition, for the subset of questions where human-vs-human IRR scores were available, the CAT generally achieved similar or higher agreement with humans than humans achieved with one another. Taken together, this indicates an overall performance comparable to human experts but demonstrates that all theme prevalence statistics should be interpreted as *estimates* to communicate uncertainty.

- **Demographic bias:** We found no evidence of systematic differences in accuracy across observed demographic groups, our proxy measure for algorithmic bias. Design features of the CAT further mitigate risk of demographic bias.

Benefits

- **Efficiency and financial impact:** If scaled across DfT's full consultation portfolio, the CAT could achieve savings in the range of £1.5-4 million per year and alleviate significant resource pressure (see Section 10).

- **All responses are always processed:** Every response is processed by the CAT for both theme generation and theme mapping. This can differ from human-analysed consultations, where a sample of responses is sometimes analysed for larger consultations.

- **Even small-scale consultations benefit:** Because there are some fixed costs to the consultation analysis process – including human-in-the-loop theme review, report writing, and project management – the cost and time savings from the CAT increase for larger consultations. However, users reported positive feedback from even small-scale (< 50 responses) consultation surveys, citing time savings and the usefulness of the structured, consistent approach of the CAT analysis.

Design and methodological learnings

- **Human-in-the-loop design:** We observed variation in the accuracy across datasets, which highlights the importance of the human review stage for theme generation (see Section 6). Based on our strong theme mapping results, and the fact that summary statistics are communicated as estimates with confidence intervals, we suggest that formally reviewing the accuracy of the CAT's subsequent theme mapping analysis to enable IRR scores to be calculated, should be an

optional step for users but is not a requirement. We conclude that human resource is best invested in a) reviewing and validating the AI-generated themes and b) synthesising the final results to shape policy decisions which involves reading responses to select illustrative quotes that represent key views.

- **Use Bayesian updating to inform the number of responses to review:** A key challenge was deciding how many responses were appropriate to review when conducting the human-in-the-loop theme review. We developed an explainable, assumption-based model to estimate the minimum number of responses required to be reviewed to be approximately 95% confident that no main themes were missed (see Annex 3 for details). This model produces a prior expectation on how many responses to review. Users can then update this prior expectation based on the qualitative feedback loop from the act of reviewing the data itself. This helps users plan, invest scarce resource most pragmatically, and make informed decisions.

- **A pure LLM methodology worked best**: The 'pure' LLM-based methodology used for CAT v1.0 outperformed the other methods evaluated that blended LLMs with more traditional machine learning approaches (see Annex 4 for details).

- **LLM ensembling boosts performance**: Using an ensemble of LLMs, where free-text data is analysed multiple times before providing a consolidated set of themes, improved theme generation performance. Similarly, using a majority-vote LLM ensemble method for theme mapping – where a theme is classified to a response only if a majority of LLMs agree on the same classification – improved performance, particularly for evaluation metrics that measured changes to the rank of themes.

- **Explainability was valued by users**: Users liked explainability features, such as providing a CAT-generated rationale for analytical steps (see Section 9).

- **Ground truth is slippery in qualitative analysis:** Language is subjective, experts legitimately disagree, and coding errors or variability creep in when analysing qualitative data. Comparisons to human-analysed datasets are therefore inherently contestable, making evaluation challenging. For example, variability in the CAT's accuracy results across datasets is partly a function of the variability in the quality and heterogeneity of the human-analysed reference datasets themselves. This renders a perfect evaluation accuracy score of 1 as effectively unachievable in many cases.

- **Communicating uncertainty in LLM-driven analysis:** Because model performance can vary across datasets and thematic contexts, we explored different ways to represent uncertainty in the LLMs' classifications in the mapping step. Initially, we tested whether statistical uncertainty derived from the model's log probabilities of output tokens during the theme mapping stage could effectively communicate variation in confidence across classifications. We concluded that presenting bootstrap confidence intervals provided a more practical and interpretable way to convey uncertainty of theme prevalence rates to users.

1. Introduction

1.1 Background

The Department for Transport (DfT) has an internal AI programme to effectively adopt and harness the potential of AI at pace. The Consultation Analysis Tool (CAT) supports commitments in the Transport AI Action Plan to use AI to "drive efficiencies in DfT's operations" and "analyse public consultation responses more rapidly and accurately".[2]

The DfT runs around 55 public consultations annually,[3] with some consultations receiving hundreds of thousands of responses. Using human researchers to convert unstructured free-text comments into structured, thematically analysed data is highly resource intensive and typically consumes over half of the entire consultation budget and can take months.

Using AI to support this process is an obvious use case. Informed by DfT's (2023)[4] research, which found that the public are receptive to AI being used to analyse consultation responses, the DfT AI and Data Science team and The Alan Turing Institute co-developed and evaluated the CAT.

This report presents our evaluation findings, which demonstrates that the CAT accurately analyses free-text consultation responses whilst unlocking efficiency and cost-saving benefits. Publishing this evaluation aims to foster public trust in the use of AI[5] and demonstrate that the DfT's use of the CAT accurately and responsibly captures the voice of consultees.

1.2 Evaluation questions

The core evaluation questions were:

---

[2] Transport artificial intelligence action plan - GOV.UK

[3] Includes consultations and calls for evidence

[4] Department for Transport. (2023). Public attitudes to the use of AI in DFT consultations and correspondence. https://www.gov.uk/government/publications/public-attitudes-to-the-use-of-ai-in-dft-consultations-and-correspondence

[5] Despite a recent survey identifying transport as an area of public services where adults would trust AI the most, only 30% of respondents agreed that they would trust the government to use AI to complete some of its tasks (ONS, 2024).

i. How accurate is the CAT compared to human-analysed datasets for theme generation and theme mapping when using a *blind evaluation design*?

ii. How accurate is the CAT when used in a live pilot setting, where human reviewers can see, review, and amend the CAT's initial analysis, and the original CAT output is then compared to the final human-validated analysis (*non-blind approach*)?

iii. To what extent is there evidence that the CAT is systematically less accurate for certain protected characteristics (our proxy for bias)?

iv. What are the learnings from the human-in-the-loop design used in our pilot of the CAT in a live consultation setting?

1.3 Context and related literature

1.3.1 Human vs AI-assisted approaches to thematic analysis

Thematic analysis is a qualitative method for identifying, analysing, and interpreting patterns (themes) in text data. The final output of thematic analysis in our context is a survey question where a set of themes have been identified and mapped to each free-text response that mention each theme. Multiple themes can be assigned to any single response. Different disciplines have employed different methodologies to achieve this.

In social research, human researchers typically first familiarise themselves with a sample of responses and apply descriptive codes to meaningful text segments, then iteratively group and refine these codes into higher-level themes. This stage relies on subjective judgment, often requiring iterative discussion and refinement to ensure that themes reflect the underlying data. This process results in a 'codebook' (a set of recurring themes) which is then used to classify (map) themes to the remaining responses to understand their relative prominence. A codebook is typically created from a sample of responses; DfT's central guidance on thematic analysis recommends 100-200 although there are published examples below and above this range. The concept of theme 'saturation' can also guide this decision, where additional responses are read and new themes are added to the codebook until a researcher is confident that there are no new themes to be identified. [6] [7]

In machine learning, unsupervised topic modelling algorithms can be used to cluster semantically similar text and surface representative topic labels. While scalable, they face several limitations compared to human-led thematic analysis. Human researchers can generate more meaningful and contextually grounded themes, drawing on detailed policy knowledge. They can also articulate their reasoning in natural language, offering greater transparency and interpretability.

Breakthroughs in LLMs offer a third way: the ability to rapidly generate contextually grounded themes. Although leveraging LLMs to conduct thematic analysis is relatively

[6] Saturation in qualitative research: exploring its conceptualization and operationalization

[7] To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales: Qualitative Research in Sport, Exercise and Health: Vol 13, No 2

simple, designing systems that align with human preferences, have an appropriate level of human oversight, and have a robust performance evaluation framework is more complex.

1.3.2 Benchmarking performance

Benchmark metrics

Performance in thematic analysis can be evaluated from two distinct perspectives.

Firstly, inter-rater reliability (IRR) metrics assess the consistency with which two or more researchers analyse the same dataset. In our context, this is typically 'labelling' or 'classifying' (mapping) the same agreed set of themes to the same free-text responses, treating both researchers as peers with no assumption that one rater's analysis is the authoritative 'true' gold standard. IRR is typically used to measure agreement between human researchers, where perfect agreement is seldom achieved due to there being subjectivity in the task (McHugh, 2012).[8] It has also been used in AI research to measure the agreement between LLMs and human researchers (Badshah & Sajjad, 2024).[9] A range of IRR measures exist, with the proportion of labels agreed upon being the most readily interpretable measure. The Cohen's kappa coefficient (for two raters) and Fleiss's kappa (> 2 raters), which ranges from -1 and +1, are common IRR statistics which aim to adjust for chance agreement and are widely used in social science literature. Cohen (1960)[10] suggested the following rough heuristics to help interpret the size of the coefficients in qualitative terms: no agreement (< 0.0), slight agreement (0-0.20), fair agreement (0.21-0.40), moderate agreement (0.41-0.60), substantial agreement (0.61-0.80), or almost perfect agreement (0.81-1.0). McHugh (2012) forwards more stringent thresholds for moderate agreement (0.60-0.79), strong agreement (0.80-0.90), and almost perfect agreement (> 0.90). However, one recognised limitation of kappa is that the results are penalised when class distributions are imbalanced which is often the case in thematic analysis; for example, when there are many more instances of theme absence (0) than theme presence (1). In addition, it assumes that there's a genuine risk that raters would randomly guess when uncertain on a particular labelling decision – an unrealistic scenario in professional settings. For these reasons, it is recommended to report the raw proportion agreement rate (the number of labels agreed upon divided by the total number of labels) alongside kappa (McHugh, 2012).

The second perspective evaluates how accurately a system (such as an LLM) compares to a human-analysed dataset, treating the human analysis as the 'true' gold standard. This approach allows the calculation of accuracy-based metrics widely used to evaluate model performance in machine learning, including recall (proportion of true themes identified), precision (proportion of predicted themes that are correct), and F_1 (harmonic mean of recall and precision). Ideally, a human-analysed reference dataset *itself* has an IRR score from two (or more) human researchers to indicate its mapping reliability and dataset subjectiveness. This approach also means that the final human-analysed gold standard

[8] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.
[9] Badshah, S., & Sajjad, H. (2024). Reference-guided verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form text. *arXiv preprint arXiv:2408.09235*.
[10] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

reference dataset can be determined by ‘majority vote’, where a label is chosen only if more than half of the analysts agree on the same label.

As explained below, establishing a benchmark and method for assessing the quality of the theme generation task is more practically challenging than the downstream task of mapping themes to responses.

Theme generation benchmarks

As noted above, using AI to generate a set of themes approximates the role of a codebook in human analysis. Developing a codebook through traditional human methods is typically iterative and very time-consuming. Notably, in the studies we reviewed we found no examples of human-only research teams independently creating separate codebooks from the same dataset and then comparing the outputs. We believe this is likely due to the inherently subjective nature of theme generation, which often requires researchers to collectively agree on an initial set of themes. This shared understanding ensures consistency in categorising the data, making independent codebook development less practical or desirable in human-led qualitative research.

Player et al. (2024)[11] developed a method to compare themes produced by humans with those generated by an LLM using four survey datasets. This involved two researchers independently triangulating LLM themes with human-generated ones to ascertain whether the themes were “in agreement” (conceptual convergence between themes), “complementary” (shared meaning of essence between themes), “dissonant” (disagreement between the themes), or “silent” (absent themes). Their DECOTA model captured 80-100% (varied by question) of human themes (recall) to an agreement or complementary level, with 10% silent themes overall. They also measured how the two researchers “broadly agreed” on the triangulation exercise necessary to calculate these statistics; the IRR scores for proportion agreement ranged between 83% and 100% whilst Cohen’s kappa ranged between 0.67 and 1.0. Scores tended to be higher in datasets with less variability and where more specific prompts were used.

Where human triangulation isn’t practical, AI approaches can be used to determine whether two pieces of text are approximately similar. One approach is to measure numerical similarity directly (e.g., cosine similarity between embedding vectors), but an increasingly common method is to develop an LLM-as-a-Judge evaluator. As the name implies, this involves prompting an LLM to makes this comparison (cf. Liu et al., 2024)[12] and it is recommended to use a human-validated method and judgment criteria. Francis et al. (2024)[13] evaluated 20 LLMs’ performance in predicting true themes generated by humans and used a validated (F_1=0.81) LLM-as-a-Judge to determine predicted-vs-true theme match pairs. They found that the percentage of true themes predicted (recall) varied

[11] Player L, Hughes R, Mitev K, Whitmarsh L, Demski C, Nash N, Papakonstantinou T, Wilson M. The use of large language models for qualitative research: The Deep Computational Text Analyser (DECOTA). Psychol Methods. 2025 Apr 7. doi: 10.1037/met0000753. Epub ahead of print. PMID: 40193412.

[12] Liu, Y., Zhou, H., Guo, Z., Shareghi, E., Vulić, I., Korhonen, A., & Collier, N. (2024). Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

[13] Francis, J., Esnaashari, S., Poletaev, A., Chakraborty, S., Hashem, Y., & Bright, J. (2024). MIMDE: Exploring the Use of Synthetic vs Human Data for Evaluating Multi-Insight Multi-Document Extraction Tasks. *arXiv preprint arXiv:2411.19689*.

significantly by model, ranging between 0.46 and 0.80. Performance was much higher on a synthetic LLM-generated dataset.

Theme mapping benchmarks

In human analysis, IRR scores are commonly calculated for the theme mapping step; it involves tasking two (or more) researchers to classify responses with a pre-defined set of themes (or codes) and then comparing their analysis. As mentioned, although it depends on the dataset, an IRR kappa score above 0.6 is typically interpreted as either substantial or moderate agreement between two human researchers so provides an established rough benchmark. Numerous studies have constructed human benchmark comparisons.

Gilardi et al. (2023)[14] conducted a study in which research assistants constructed a gold standard dataset by assigning qualitative categories to text.[15] They then performed these exact same classifications with GPT 3.5 turbo zero-shot prompting and with crowd-workers recruited on MTurk, using the same codebook developed by the research assistants. ChatGPT outperformed crowd-workers for several annotation tasks. Proportion agreement IRR scores (percentage of instances for which both annotators *within* a given group report the same class) were also calculated. This was approximately 56% for MTurk, 79% for trained annotators, 97% for ChatGPT with a low temperature (0.2).[16]

Kirsten et al. (2024)[17] evaluated the performance of OpenAI models and humans to annotate a dataset using the same human-generated codebook. They found that IRR kappa scores for human researchers ranged between 0.63 and 0.97, with a deterioration in performance for more complex datasets with longer codebooks. Although they found that GPT-4 performed better than GPT-3.5, it performed worse on the more complex tasks (0.56) compared to simpler ones (0.97). Using a temperature of 0, they found that few-shot prompting (i.e., providing the model with examples) did not improve IRR scores but reduced hallucinations. They also concluded that future research should consider how to communicate error and that the human benchmark itself is not infallible, as human researchers are susceptible to bias and subjectivity.

Makinson et al. (2025)[18] undertook an evaluation of an AI consultation analysis tool with a similar purpose to the CAT. For their evaluation, they used an AI to map themes to responses and then a human team reviewed the mapping and, where necessary, amended the themes that the AI had mapped to each response. Reviewers could also add new themes to a response at this stage and this human-altered dataset acted as the reference dataset in the evaluation. They achieved an average F_1 of 0.76 across all questions and reviewers made no changes to 60% of responses. In another evaluation, they achieved mappings that were highly consistent with human reviewers, achieving F_1

---

[14] Fabrizio Gilardi, Meysam Alizadeh, & Maël Kubli. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(30). https://doi.org/10.1073/pnas.2305016120

[15] By gold standard we mean a curated dataset of responses with a high-quality code-frame and well-mapped themes.

[16] See definition of temperature here: https://docs.claude.com/en/docs/about-claude/glossary

[17] Elisabeth Kirsten, Annalina Buckmann, Abraham Mhaidli, & Steffen Becker. (2024). Decoding Complexity: Exploring Human-AI Concordance in Qualitative Coding. *ArXiv.Org*. https://doi.org/10.48550/arxiv.2403.06607

[18] https://ai.gov.uk/docs/Scot_Gov_NSCP_Evaluation_Report.pdf

scores of 0.79 and 0.82, higher than the agreement between human reviewer groups themselves. Reviewers added new themes for around 12% of responses.[19]

Francis et al. (2024) evaluated a scenario where an LLM-mapped dataset was compared to a human-mapped dataset. In contrast to the above studies, the LLM-mapped themes were generated by an LLM whereas the reference human-mapped dataset used human-generated themes. The LLM performed comparatively poorly in this scenario (max F_1=0.45), an intuitive finding as whenever a theme was missed in the initial LLM theme generation stage, it could not be correctly mapped to responses, leading to compounded errors.

1.3.3 Demographic bias

In the context of AI, demographic bias is the systematic difference in the treatment of certain groups of people. A fair algorithmic system seeks to mitigate any such prejudice/favouritism in its use.[20]

LLMs are typically trained in two main stages: first on vast amounts of curated data (pretraining), then fine-tuned using alignment techniques such as reinforcement learning from human feedback (RLHF) and other methods to help make their outputs helpful, safe, and aligned with human values. LLMs can therefore reflect stereotypes, prejudiced beliefs, language, and predispositions of humans that generated the data initially, so there is legitimate concern that LLMs may learn and amplify some of these social biases, although this risk depends on the specific application.[21] In addition, LLMs can perform worse when analysing language that is underrepresented in the training data, known as an 'out-of-distribution' limitation.

The main channel through which systematic demographic bias could occur in our context, given that the CAT does not explicitly use demographic variables within any of the LLM prompts, is the well-documented channel of language differences (Reusens et al., 2023; Gallegos et al., 2024).[22] [23] For example, an LLM may perform worse on responses that are written in poor English or use socio-culturally specific language (e.g., verbosity, slang). To the extent that this can be correlated with demographics – e.g., use of specific slang could be correlated with age – performance could be worse for certain demographics on average. Work by Ashwin et al. (2025) found in their study that the prediction errors LLMs "make in coding are not random with respect to the characteristics of the interview subject", but did not find a consistent pattern across demographics and across different codes and models.[24] This "can be problematic if these biases and errors systematically correlate with people's characteristics, such as gender, education, race and ethnicity, and

[19] https://ai.gov.uk/evaluations/consult-evaluation-independent-water-commission-call-for-evidence/

[20] ISO/IEC TR 24027:2021 - Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making

[21] Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R. and Ahmed, N.K., 2024. Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), pp.1097-1179. https://doi.org/10.1162/coli_a_00524

[22] Ibid

[23] Reusens, M., Borchert, P., De Weerdt, J., & Baesens, B. (2024). Native design bias: Studying the impact of english nativeness on language model performance. *arXiv preprint arXiv:2406.17385*.

[24] Ashwin, J., Chhabra, A., & Rao, V. (2025). Using Large Language Models for Qualitative Analysis can Introduce Serious Bias. *Sociological Methods & Research*, *0*(0). https://doi.org/10.1177/00491241251338246

socioeconomic status".[25] It is important to emphasise that the same biases can exist for human analysis too; in DfT's (2023)[26] research, participants raised concerns that consultation analysis using AI could be less accurate on responses written in "poor English" but also reasoned the same applied to human analysis.

We assessed whether we could identify systematic bias empirically (see Section 8).

1.3.4 Human oversight

There are numerous defensible approaches to incorporating human oversight into AI-assisted thematic analysis.

One approach is to use human researchers to generate a set of themes before using an LLM to map themes to responses (cf. Kirsten et al., 2024). DHSC (2024)[27] used topic modelling in combination with officials reviewing a sample of representative quotes to construct theme labels for each topic. Ofqual (2024)[28] chose to validate LLM-generated themes by reviewing the 20 most semantically similar sentences to each extracted theme but did not subsequently map themes to responses. Makinson et al. (2025)[29] used a similar approach at the theme generation stage, where themes were presented to policy professionals and example responses for each theme were shown. They also had humans verify and, if necessary, amend the mapping of themes to responses.

Consultation survey responses are not statistically representative of a given population as they're generally self-selecting. Therefore, the prevalence of a theme cannot be interpreted as the proportion of the wider public holding a particular view and a higher prevalence theme is not necessarily qualitatively more influential to a policy decision. DfT internal guidance on analysing consultations acknowledges this and, noting that coding error can occur in human analysis, encourages themes to be qualitatively interpreted as broad proportions rather than exact percentages e.g., "a dominant view was" rather than "47% of respondents felt". Based on this, we decided it was particularly important to invest scarce human resource into assuring the accuracy and quality of the theme generation step – equivalent to creating a codebook in traditional social research – compared to the mapping step. This design choice accepts some degree of error in the process of mapping themes to responses and assumes that the value of findings from qualitative data predominantly lies in depth, context, and interpretation. We believe our mapping evaluation results (Section 7) and use of confidence intervals to convey uncertainty in percentages justifies this choice, but CAT users can (optionally) complete a formal review of the mapping step to calculate IRR scores. There is also human oversight during the synthesis step to interpret the analysed data when writing the report, which typically involves reading responses to extract representative quotations.

---

[25] Liu, A., & Sun, M. (2025). From Voices to Validity: Leveraging Large Language Models (LLMs) for Textual Analysis of Policy Stakeholder Interviews. *AERA Open*, *11*. https://doi.org/10.1177/23328584251374595

[26] Department for Transport. (2023). Public attitudes to the use of AI in DFT consultations and correspondence. https://www.gov.uk/government/publications/public-attitudes-to-the-use-of-ai-in-dft-consultations-and-correspondence

[27] Creating a smokefree generation and tackling youth vaping consultation: government response - GOV.UK

[28] Findings of Ofsted's Big Listen public consultation - GOV.UK

[29] Scot Gov NSCP Evaluation Report

2. Ethics

2.1 Summary

We have endeavoured to align the CAT with the ethical principles set out in the government's AI Assurance framework.[30]

2.2 Transparency

This publication supports transparency and DfT has been transparent about piloting use of AI since this work began in 2023. Firstly, the DfT has clear privacy notices on use of AI in processing consultation surveys[31] and personal information charter.[32] The DfT has also referred to this work in parliament,[33] the Transport AI action plan,[34] and externally at events including the DfT Data Science in Transport Conference[35] and DfT-Google-PA AI Hackathon in Transport.[36] In addition, an Algorithmic Transparency Recording Standard (ATRS)[37] document has been drafted for the CAT and its publication is forthcoming. The ATRS provides clear information about public sector use of algorithmic tools, including the rationale for their use, the responsible officials and how the model works.

2.3 Data protection

A Data Protection Impact Assessment (DPIA) for the CAT is in place and remains under review. Each consultation that makes use of the CAT is subject to data protection compliance measures.

2.4 Consideration of public attitudes

The legitimacy and effectiveness of the consultation process rely significantly on public trust in the methods used. DfT conducted specific research to understand public

[30] Introduction to AI assurance - GOV.UK
[31] DfT online forms, surveys and consultations privacy notice - GOV.UK
[32] Personal information charter - Department for Transport - GOV.UK
[33] Q. Department for Transport: Artificial Intelligence
[34] Transport artificial intelligence action plan - GOV.UK
[35] Data Science in Transport Conference
[36] AI Hackathon in Transport in collaboration with DFT and Google Cloud
[37] Algorithmic Transparency Recording Standard Hub - GOV.UK

perceptions and concerns regarding the use of AI for analysing consultation responses. The findings were generally positive, indicating a level of comfort with the proposed application: *"Overall, both the interviews and workshops showed that the public generally felt comfortable with the use of AI to analyse consultation responses".*[38]

2.5 Human oversight by design

Maintaining meaningful human oversight, or a human-in-the-loop, is a core ethical safeguard in AI deployment. This is designed to ensure potential AI errors or misinterpretations are identified, and to keep human judgment central to understanding public input.

There are many ways in which human oversight can form part of an AI process. Our approach formally integrates human oversight in the theme review step (see Section 3) and at the analysis and report writing stage, where users interrogate the CAT-enabled analysis and select representative quotations. The division of labour between AI and humans was found to resonate positively in the public research we completed as we started developing the CAT: "*participants were generally satisfied with the level of human oversight for this (AI for consultations) use case. The fact that AI would perform the data analysis but humans would still be responsible for crafting the final consultation response, meant that the process still allowed humans to perform a critically authoritative part of the process".*[39]

2.6 Bias mitigation

We have considered risks of unintended demographic bias when designing and developing the CAT. In Section 8, we present empirical analysis which considers whether performance systematically varies by protected characteristic.

2.7 Accountability

The human decision-maker remains accountable for any policy decisions that follow a consultation.

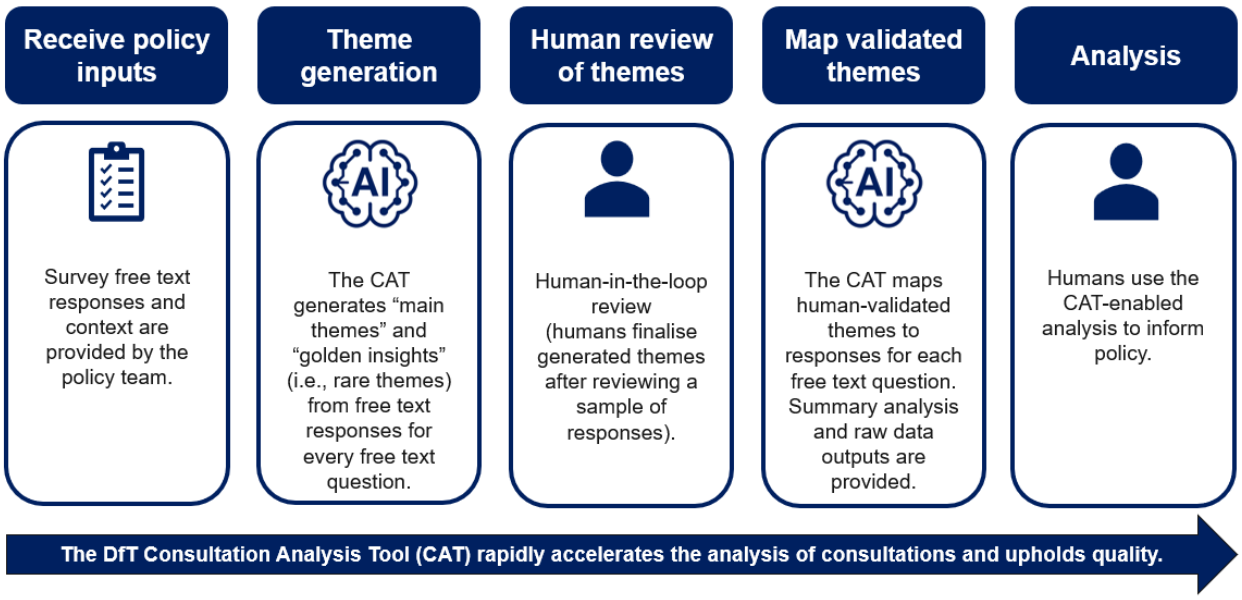
[38] Public attitudes to the use of AI in DfT consultations and correspondence - GOV.UK
[39] Ibid

3. CAT implementation methodology

3.1 Overview

The below schematic summarises the key steps of the CAT v1.0 methodology and process, which was decided after assessing the alternative methodologies outlined in Annex 4. For CAT v1.0, we used the Google Gemini family of LLMs.[40]

Figure 1 Overview of the CAT process with human-in-the-loop theme review



3.2 Theme generation step

The theme generation process involves the following key steps:

[40] Other models were tested and evaluated at different stages of developing the CAT. Prompts had to be re-engineered and evaluated when we changed and upgraded models during the CAT's development.

- **Rich data input and contextual understanding**: The CAT is provided with consultation context and specific question details, alongside the free-text responses, required to accurately interpret and analyse the responses.[41]
- **Multiple analysts**: The system makes multiple LLM calls (each representing a different 'analyst') to independently identify themes from the same dataset. For each analyst, responses are randomly shuffled to mitigate positional bias and responses are batched.[42] This leads to significant duplication of themes across analysts and batches but results in a more diverse set of themes.
- **Theme consolidation**: After generating themes from multiple analysts, another LLM pipeline merges similar themes to create a cohesive final set. Each theme has a title and a description.
- **Sub-group analysis (optional):** The CAT (depending on user choice) can extract separate theme sets by different groups, such as multiple-choice questions (e.g., agree vs disagree) and/or respondent type (e.g., organisation vs member of the public), if this additional context is deemed important for accurately extracting themes.

Alongside the central feature of main theme generation, key features include:

- **Golden insights extraction**: Golden insights (i.e., rare themes) are extracted, which may have only been mentioned by a few respondents but nonetheless have high policy relevance.
- **Explainability and auditability:** Users are presented with an audit trail for each stage of the theme generation process, where they can read an explanatory comment justifying why the AI chose to generate and consolidate each particular theme.

3.3 Human-in-the-loop theme review step

The CAT v1.0 methodology used in our live pilots incorporated a human-in-the-loop theme review phase following the automated AI theme extraction step. After the CAT generated an initial set of themes, human reviewers examined a random sample of responses to gain direct understanding of the consultation responses to assess the quality of the themes. Users were provided with guidance to decide on the number of responses to be manually reviewed at this stage, which combined the use of a quantitative model and social research literature (see Annex 3). The human-in-the-loop review process took between 1 and 5 hours per 100 responses (varied by dataset).

At this stage, reviewers interrogated the CAT-extracted themes and could make refinements such as:

- Removing any themes deemed unimportant from a policy perspective.

---

[41] Rich context, such as defining key terms for the consultation, is important here for addressing possible "out-of-distribution" LLM limitations required to interpret language underrepresented in training data.

[42] Cf. Lui et al. (2023) for research on how LLM performance can degrade in long contexts.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv preprint* arXiv:2307.03172.

- Adjusting the wording of an existing theme.
- Merging similar themes or disaggregating themes.
- Adding new themes that the CAT may have missed but are evident in the responses.
- Confirming all validated themes are accurately representative of the responses.

This process results in a final human-validated set of main themes. Golden insights are also reviewed and, where necessary, refined or dropped at this stage too. This human oversight ensures that the final set of themes reflect both the systematic analysis of the CAT and the judgment of experienced reviewers.

This curated set of main themes and golden insights comprises the human-validated input to the response-level mapping phase.

3.4 Response level theme mapping (multi-label classification) step

The CAT, using an LLM pipeline, systematically analyses each individual response to identify which of the human-validated themes are mentioned. This mapping process, which uses a majority-vote ensemble of LLMs approach – where a theme is classified to a response only if a majority of LLMs agree on the same classification – creates a comprehensive mapping between responses and themes. An ensemble was selected for performance reasons and bias mitigation (Badshah & Sajjad, 2024).[43]

For each response, the CAT determines:

- Which of the identified themes are represented in the response.
- Which response(s) represent the golden insights (i.e., rare themes).

The system also has an optional step to classify responses that deviate from common patterns.[44]

The outcome is a rich, structured dataset that preserves the depth of qualitative insights while enabling the analytical power of quantitative methods. To communicate uncertainty in the estimated prevalence of each theme, statistics with bootstrap-derived confidence intervals are presented to users. This involves repeatedly resampling (with replacement) from the analysed dataset 1,000 times and recalculating theme prevalence for each resample. The 2.5th and 97.5th percentiles of these estimates are then taken to form a 95% confidence interval for each theme's prevalence. The intervals capture sampling uncertainty, that is, how much the estimated prevalence would vary if a different but comparable sample of responses were analysed. Users are also made aware of the evaluation results to contextualise model reliability.

---

[43] Badshah, S., & Sajjad, H. (2024). Reference-guided verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form text. *arXiv preprint arXiv:2408.09235*.

[44] The CAT flags responses for special attention based on criteria like irrelevance, incoherence, or offensive content. This outlier detection is customisable to highlight categories of interest such as concrete proposals, personal experiences, disability perspectives, or technical expertise.

4. Evaluation data sources

4.1 Performance evaluation data overview

As described in the Executive Summary, our performance evaluation compared the CAT analysis to human-analysed datasets. The same free-text survey responses were analysed by humans and the CAT, resulting in (true) human-analysed outputs and (predicted) CAT-analysed outputs. We then compared the CAT analysis to the human-analysed reference data, enabling our performance metrics to be calculated (Section 5). We use the terms 'blind' and 'non-blind' to clearly distinguish between how the following different datasets were produced and analysed (see Annex 1 and Annex 2 for details).

1) **Datasets used in the *blind evaluation design*:** This covered 11 questions, 9,100 responses, and 165 main human-validated 'true' themes across three overall datasets. These overall datasets are named INTS (3 questions), ATI_A (5 questions), and ATI_B (3 questions). The INTS human-analysed questions were rigorously produced by two researchers at DfT and the others were created by researchers at The Alan Turing Institute with support from crowdsourced workers (see Annex 1). For this blind evaluation design, the CAT analysed the same survey data as humans and we then compared the two sets of analysis to calculate our performance evaluation metrics. The human-analysed datasets were therefore created without the risk of anchoring bias influencing the analysis.[45]

2) **Datasets used in the live pilot *non-blind evaluation design***: This covered 32 questions, 198,000 responses, 480 'true' main themes, and 749 'true' golden insights across four overall datasets: DVSA (6 questions), INTS (19 questions), NZP (6 questions), and UKR (1 question). The questions were from DfT consultations; 32 were used for our theme generation evaluation and seven (total responses=720) were used for our theme mapping evaluation (see Annex 2 for precise details). For this non-blind evaluation design, the CAT-generated themes were used as a starting point, and human reviewers then modified items they disagreed with after analysing the data themselves. Although the human-validated reference data faces the risk of anchoring bias, it benefits from modifications being completed by those familiar with the dataset and policy area of the consultation.

[45] Anchoring bias is the human tendency to rely on an initial value (an 'anchor') when making subsequent judgements. We acknowledge a risk of *indirect* overfitting to these evaluation datasets through iterative development, although the diversity of our datasets helps to mitigate this risk.

5. Performance evaluation metrics

5.1 Theme generation performance metrics and procedure

We developed typical classification metrics that compared the set of CAT-generated (predicted) themes to the set of human-generated (true) themes.[46]

- **Recall (lead metric of interest):** The proportion of the true themes that have been identified.[47]
- **Precision:** The proportion of the predicted themes matched to the true themes.
- **F_1:** The harmonic mean of recall and precision.

The above metrics require a triangulation exercise to assess, theme-by-theme, which CAT-generated themes match the human-generated theme set, which is a resource-intensive task. We used different approaches for each evaluation design.

Our first approach – used in our automated *blind evaluation design* results (see Figure 2) – uses an LLM-as-a-Judge framework to complete the triangulation exercise.[48] This involved an LLM being prompted to judge whether each CAT (predicted) and human (true) theme pair was approximately the same.[49] This triangulation exercise is a similar approach to Player et al. (2024),[50] but we used an LLM-as-a-Judge instead of human researchers to judge whether a theme was 'the same'. Importantly, the LLM-as-a-Judge was *itself* evaluated on a rich labelled dataset, which achieved a macro F_1 of 0.8. This work built on the method and labelled data published by Francis et al. (2024).[51] Similarly to Shankar et

[46] These metrics are calculated by comparing two overall sets of themes (not at the response level).
[47] Percent of true themes 'missed' by the CAT is the complement of recall (1 minus recall).
[48] Theme triangulation here uses a Cartesian product between the predicted and true themes. For example, if there are 12 predicted themes and 10 true themes, this yields 12 × 10 = 120 pairwise comparisons.
[49] Themes were judged as the same when they express the same insight, even if phrased differently. We developed a specific rubric for this task.
[50] Player L, Hughes R, Mitev K, Whitmarsh L, Demski C, Nash N, Papakonstantinou T, Wilson M. The use of large language models for qualitative research: The Deep Computational Text Analyser (DECOTA). Psychol Methods. 2025 Apr 7. doi: 10.1037/met0000753. Epub ahead of print. PMID: 40193412.
[51] Francis, J., Esnaashari, S., Poletaev, A., Chakraborty, S., Hashem, Y., & Bright, J. (2024). MIMDE: Exploring the Use of Synthetic vs Human Data for Evaluating Multi-Insight Multi-Document Extraction Tasks. *arXiv preprint* arXiv:2411.19689.

al. (2024),[52] we also conducted iterative reviews of the judgements to align the LLM-as-a-Judge decisions with human preferences through refinements to prompts and LLM specification. We then used this validated judge for the triangulation decisions in our automated theme generation evaluation pipeline.

Our second approach – used for our *non-blind evaluation design* – involved comparing a human-validated set of themes to the initial CAT-generated themes. The human-validated data was produced by the formal human-in-the-loop step, where human reviewers conducted a structured assessment of the predicted themes alongside reviewing a random sample of responses. During this review they could adjust (drop, validate, refine, and augment) themes, resulting in a human-validated set of themes.

5.2 Theme mapping performance metrics

The output from the human-mapped data is a structured sparse dataset with binary categories, indicating the presence (1) or absence (0) of each theme for each response. There are established machine learning evaluation metrics for multi-label classification problems, which are characterised as having multiple labels (i.e., themes) for each survey response. This allows us to compare the CAT mapping to human-mapped data to calculate the following metrics:

- **Multi-label recall:** The proportion of the true themes that have been identified, averaged across all responses.
- **Multi-label precision**: The proportion of predicted themes matched to true themes, averaged across all responses.
- **Multi-label F_1 (lead metric):** The harmonic mean of multi-label recall and precision.
- **Theme rank correlation:** Spearman's rank correlation coefficient between CAT-generated and human-generated theme prevalence rankings, where themes are ranked by the number of responses in which they appear.
- **Overall proportion agreement IRR**: The proportion of agreed classifications (theme absence and theme presence) between the CAT and humans.
- **Cohen's kappa score IRR**: The proportion of agreed classifications (theme absence and theme presence), adjusted for random chance agreement.

The multi-label metrics above are the average (mean) of response-level metrics, where each metric is first calculated for each response and then averaged across all responses.[53]

---

[52] Shankar, S., Zamfirescu-Pereira, J. D., Hartmann, B., Parameswaran, A., & Arawjo, I. (2024, October). Who validates the validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (pp. 1-14). https://dl.acm.org/doi/pdf/10.1145/3654777.3676450

[53] See blog by Mustufa Murat ARAT for an overview of multi-label classification metrics: https://mmuratarat.github.io/2020-01-25/multilabel_classification_metrics. We compute multi-label metrics using scikit-learn with average="samples" and set zero_division=1, which assigns a value of 1 when per-response metrics are undefined due to a zero denominator (e.g., when a response contains no true theme labels). For example, responses that are *correctly* predicted to have no themes (true negative) receive a per-response score of 1. As such cases are rare in our evaluation data, results are insensitive to this choice relative to excluding undefined instances (zero_division=np.nan): overall F_1 is unchanged in the blind setting (Fig. 6) and decreases by 1 percentage point in the non-blind setting (Fig. 8).

6. Theme generation performance results

6.1 Summary

The theme generation recall results show that, on average, the CAT identifies between 75% (blind setting) and 90% (live setting) of the same themes as human experts automatically. The human-in-the-loop theme review process ensures that the probability of extracting all 'true' main themes within the dataset approaches 100% with human review, which is how the CAT is used in practice.

6.2 Automated (blind) theme generation results

For the automated (blind) CAT evaluation design, the CAT-generated (predicted) theme set is compared to a human-generated (true) theme set. As explained in Section 5, we used an (evaluated) LLM-as-a-Judge to decide whether a predicted CAT theme matched a true theme. LLMs produce stochastic outputs, meaning results aren't fully reproducible each time the analysis is run,[54] so all results were averaged across five runs of the automated theme generation pipeline. We present bootstrap confidence intervals to illustrate the between-run variation in Figure 3, which captures variation in both the theme generation accuracy and the LLM-as-a-Judge triangulation process.

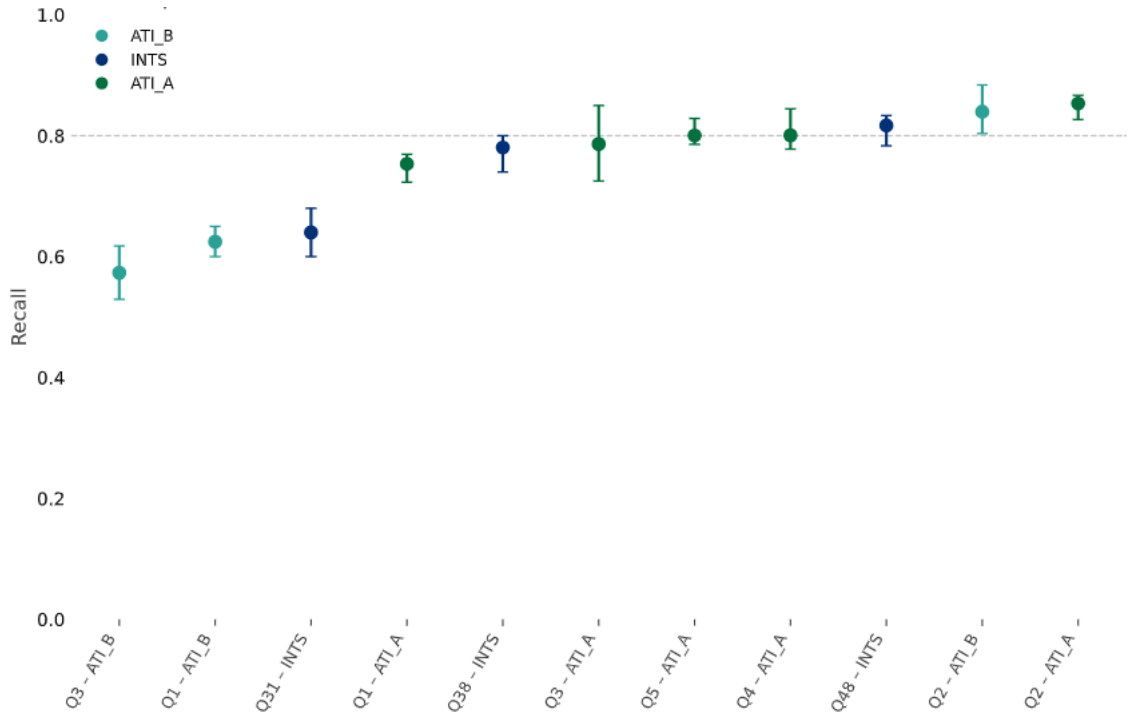
The CAT identified 75% of the themes that human researchers identified overall (see Figure 2), which is our lead metric of interest here (recall). This equates to identifying 124 out of the 165 human-generated themes across all questions. Recall varied by question, ranging from 0.57 to 0.85. Recall is the main metric of interest here as we were most concerned with detecting the same themes that humans had. The precision metric captures the proportion of the predicted themes matched to the true themes, so any themes predicted by the CAT but not matched to the true set depress precision (and F_1). The relatively low overall precision result (0.5) is therefore not necessarily 'bad' as it means that the CAT has identified themes that were missed by human analysts (a reference data quality issue), exist but not deemed a key theme by the human analysts, and/or caused by LLM hallucinations. We do not distinguish between these explanations as the human-CAT theme matches were automatically decided using LLM-as-a-Judge.

[54] There are several LLM pipeline stages during theme generation as well as the LLM-as-a-Judge step here.

Figure 2 Theme generation blind evaluation results using LLM-as-a-Judge for theme matches

	Recall	Precision	F_1
Overall	**0.75**	**0.50**	**0.59**
INTS_Q31	0.64	0.54	0.58
INTS_Q38	0.78	0.67	0.72
INTS_Q48	0.82	0.58	0.67
ATI_A_Q1	0.75	0.40	0.52
ATI_A_Q2	0.85	0.46	0.60
ATI_A_Q3	0.79	0.32	0.46
ATI_A_Q4	0.80	0.38	0.51
ATI_A_Q5	0.80	0.43	0.56
ATI_B_Q1	0.62	0.67	0.65
ATI_B_Q2	0.84	0.50	0.62
ATI_B_Q3	0.57	0.55	0.56

Figure 3 Theme generation blind evaluation results (recall) using LLM-as-a-Judge for theme matches



6.3 Live pilot (non-blind) theme generation results

The live pilot results in this section compare the initial set of CAT themes (predicted themes) to the final human-validated theme set following the formal human-in-the-loop theme review (true themes). CAT users were required to systematically review a random sample of responses alongside main themes and golden insights (i.e., rare themes),

justifying each final human-validated theme with a comment. A theme could be left unchanged (true positive),[55] substantially reworded or removed (false positive), or an entirely new theme added (false negative). We employed an arguably overly cautious definition of 'substantially reworded' here; for example, if a theme's title changed whilst the theme description remained unchanged, we still counted this as a false positive.

The overall recall, precision, and F_1 results were similar. The overall F_1 across all thirty-two questions was 0.90 (main themes) and 0.89 (golden insights) and the recall results illustrate that 90% of true main themes (n=480) and golden insights (n=749) were correctly identified by the CAT. Figure 5 illustrates variation by question, with F_1 ranging from 0.61 to 1 for main themes. The variation for golden insights (not shown) was similar. There were model upgrades and some wider methodological improvements implemented for DVSA/NZP, which likely explain the relatively higher performance on these datasets.

Users' qualitative justifications for changes made during the theme review included: themes needing to be disaggregated into two themes, leading to new themes and significant rewording; novel themes being added; streamlining the themes by merging or deleting themes that were redundant, overlapping, or duplicated; refining themes by changing titles and descriptions for conceptual clarity and emphasis; or removing them entirely if deemed irrelevant or too specific.

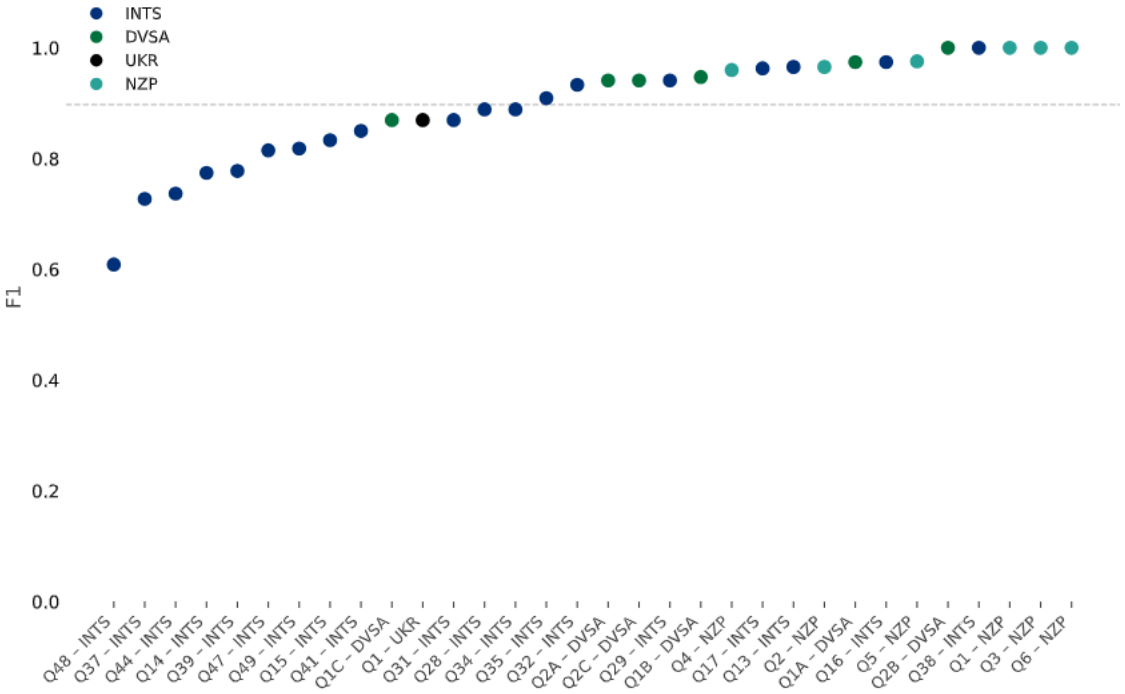
Figure 4 Theme generation live pilot (non-blind) evaluation results by data source

	Main themes			Golden insights		
	F_1	Precision	Recall	F_1	Precision	Recall
Overall	**0.90**	**0.91**	**0.90**	**0.89**	**0.92**	**0.90**
DVSA	0.95	0.90	1.00	0.98	0.97	1.00
INTS	0.86	0.89	0.84	0.82	0.88	0.84
NZP	0.98	0.97	1.00	1.00	1.00	1.00
UKR	0.87	1.00	0.77	-	-	-

Table notes: *Golden insights were generated but not shown for UKR as they're not comparable with the other data sources. UKR was the first live pilot and golden insights were generated after the theme review stage so didn't go through the same review process as the other data sources.*

[55] Very minor changes to wording e.g., using initialisms rather than full words are defined as 'the same' here.

Figure 5 Main theme generation live pilot (non-blind) evaluation results (F_1) by question



7. Theme mapping (multi-label classification) performance results

7.1 Summary

The theme mapping results demonstrate that the CAT's mapping analysis is highly aligned with human experts.

When mapping themes to individual free-text responses (i.e., multi-label classification), where the CAT analysis was compared to a human-analysed dataset using our *blind evaluation design*, the CAT achieved an overall multi-label F_1 score of 0.75 (out of 1) across 11 questions (Figure 6). During live pilots, social researchers reviewed the CAT's theme mapping analysis and changed any classifications they disagreed with. Using this *non-blind evaluation design*, comparing the initial CAT vs human-adjusted mapping, the CAT achieved an overall multi-label F_1 score of 0.93 (seven questions; Figure 8).

As described in Section 1.3, IRRs can also be used to baseline 'good' agreement rates between analysts. Figures 6 and 8 respectively show that the overall proportion of agreed classifications between the CAT mapping and human-analysed mapping was between 93% (blind setting) and 98% (non-blind setting) whilst the overall Cohen's kappa was between 0.71 (blind setting) and 0.91 (non-blind setting). Interpreting these kappa IRR results in rough qualitative terms, as suggested by others, evidence "substantial" and "almost perfect" agreement (Cohen, 1960) or "moderate" and "almost perfect" agreement (McHugh, 2012).[56] For a random sample of responses on a subset of questions, we compared CAT-vs-human results to human-vs-human IRR baselines on the same datasets. The CAT generally achieved similar or higher IRR with humans compared to IRR between two or more human researchers analysing the same questions (Annex 5).

7.2 Automated (blind) theme mapping results

The results in this section compare the CAT mapping (predicted) to the human-analysed (true) mapping data. In this setting, the themes themselves were first generated by human researchers. The average multi-label F_1 score across all questions was 0.75. Figures 6 and 7 illustrate variation by question, where F_1 results ranged from 0.62 to 0.90. The imbalance between recall and precision shows that the CAT captured most of the themes

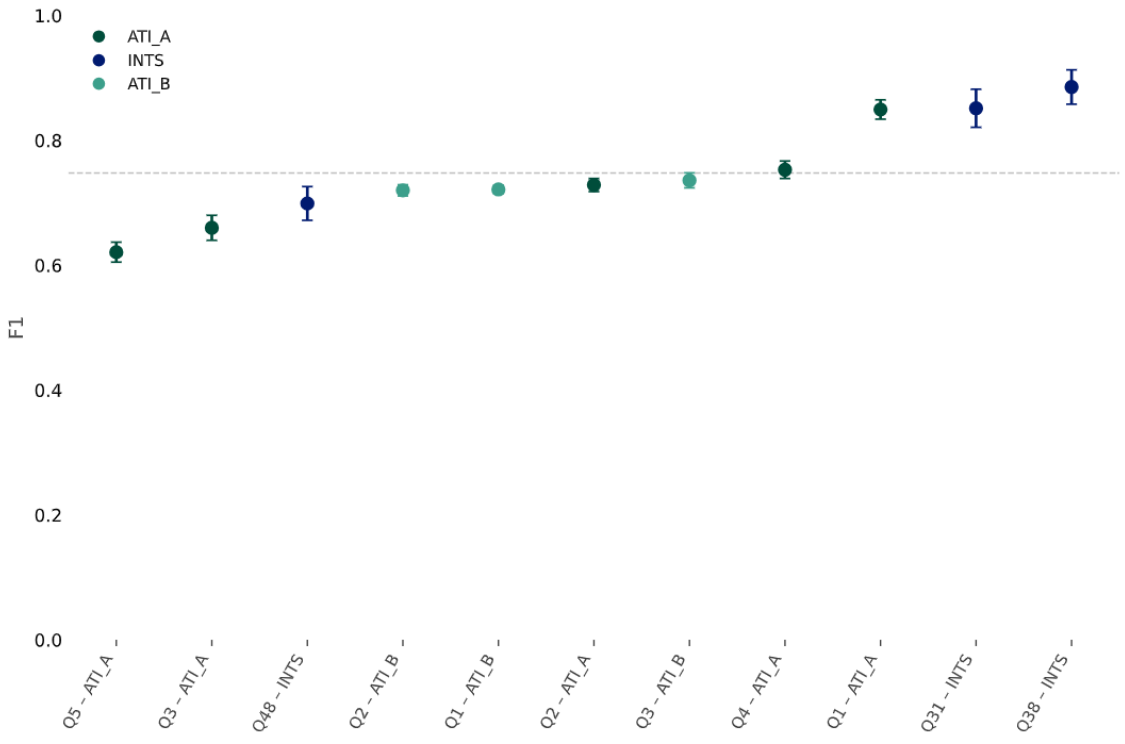
[56] Ibid.

that were identified by humans (high recall) at the cost of assigning additional themes that were not identified by humans (lower precision). The high rank correlation values indicate strong alignment between the CAT and human analysts in the relative ordering of themes by prevalence. This metric is sensitive to near ties in ranks; for example, the lower coefficient for INTS_Q48 reflects a distribution in which two-thirds of themes had a very similar number of (true) themes (± 4%), meaning that small differences in theme prevalence produced disproportionately large changes in rank and, consequently, the correlation coefficient.

Figure 6 Theme mapping blind evaluation results by question

Dataset	Number of responses	F_1	Precision	Recall	Proportion agreement IRR	Cohen's kappa IRR	Rank correlation
Overall	**9,098**	**0.75**	**0.71**	**0.90**	**0.93**	**0.71**	**0.86**
INTS_Q31	200	0.86	0.85	0.93	0.94	0.83	0.98
INTS_Q38	200	0.90	0.92	0.94	0.97	0.90	0.98
INTS_Q48	200	0.71	0.69	0.79	0.89	0.60	0.51
ATI_A_Q1	995	0.86	0.84	0.95	0.96	0.83	0.91
ATI_A_Q2	992	0.73	0.65	0.93	0.94	0.70	0.96
ATI_A_Q3	950	0.66	0.61	0.87	0.93	0.65	0.96
ATI_A_Q4	959	0.75	0.70	0.93	0.91	0.70	0.65
ATI_A_Q5	990	0.62	0.56	0.86	0.91	0.59	0.87
ATI_B_Q1	1,204	0.72	0.65	0.93	0.93	0.67	0.80
ATI_B_Q2	1,204	0.72	0.65	0.91	0.95	0.70	0.94
ATI_B_Q3	1,204	0.73	0.72	0.85	0.93	0.68	0.89

Figure 7 Theme mapping blind evaluation results (F_1) by question



7.3 Live pilot (non-blind) theme mapping results

Social researchers reviewed a random sample of the CAT mapping across seven questions (n=720), where human-validated themes (following the human-in-the-loop theme review) were mapped to responses. They then changed any classifications that they disagreed with for seven questions in our live pilot setting. The initial CAT mapping (predicted) was compared to the human-adjusted (true) dataset, achieving a very high overall multi-label F_1 score (0.93). Recall and precision were relatively evenly balanced on these datasets.

Human researchers also provided feedback on reasons that they disagreed with the CAT on certain instances. There were some examples identified where the CAT was deemed to "*read too much between the lines*" by classifying responses with related themes but without sufficient *direct* evidence in the response. Such instances depress the precision statistic. In addition, there were some edge examples where the CAT was deemed to have got "*the wrong end of the stick*", such as misinterpreting the sentiment of a response.

Figure 8 Theme mapping live pilot (non-blind) evaluation results by question

	Number of responses (sample)	F1	Precision	Recall	Proportion agreement IRR	Cohen's kappa IRR
Overall	**720**	**0.93**	**0.96**	**0.93**	**0.98**	**0.91**
DVSA_Q1A	100	0.95	0.97	0.93	0.99	0.90
DVSA_Q1B	100	0.91	0.95	0.90	0.98	0.89
DVSA_Q1C	100	0.94	0.95	0.96	0.99	0.90
DVSA_Q2A	100	0.92	0.96	0.90	0.98	0.90
DVSA_Q2B	100	0.93	0.95	0.93	0.99	0.92
DVSA_Q2C	100	0.96	0.99	0.95	0.99	0.95
UKR_Q1	120	0.94	0.98	0.94	0.98	0.92

8. Demographic bias analysis

8.1 Summary

We empirically assessed whether the CAT was systematically less accurate at mapping themes to responses for specific demographic groups. Reassuringly, we found no evidence of systematic bias on the three questions we analysed. Although this analysis is not exhaustive and faces limitations, the CAT design itself includes several safeguards to mitigate bias, including exclusion of demographic variables from prompts and the human-in-the-loop review of all CAT-generated themes. Whilst we cannot rule out demographic bias, these measures substantially reduce the likelihood of systematic bias affecting the performance of the CAT.

8.2 Demographic bias methodology

We compared CAT-analysed theme mappings (predicted) with human-analysed mappings (true), using the same set of human-identified themes (i.e., a common label space in both cases), to assess whether the CAT's accuracy is systematically lower for particular observable demographic characteristics. This analysis uses the ATI_B dataset, for which demographic information was collected for three questions (n=3,612; see Annex 1 for details).

For each respondent-question pair, we calculated a multi-label F_1 score (range 0-1), yielding one F_1 outcome per respondent per question. The resulting outcome y_i was regressed on respondent demographics and question fixed effects using the following linear specification:

$$\textbf{(1.0)} \quad y_i = \alpha + \mathbf{X}_i\beta + \delta_{q(i)} + \varepsilon_i$$

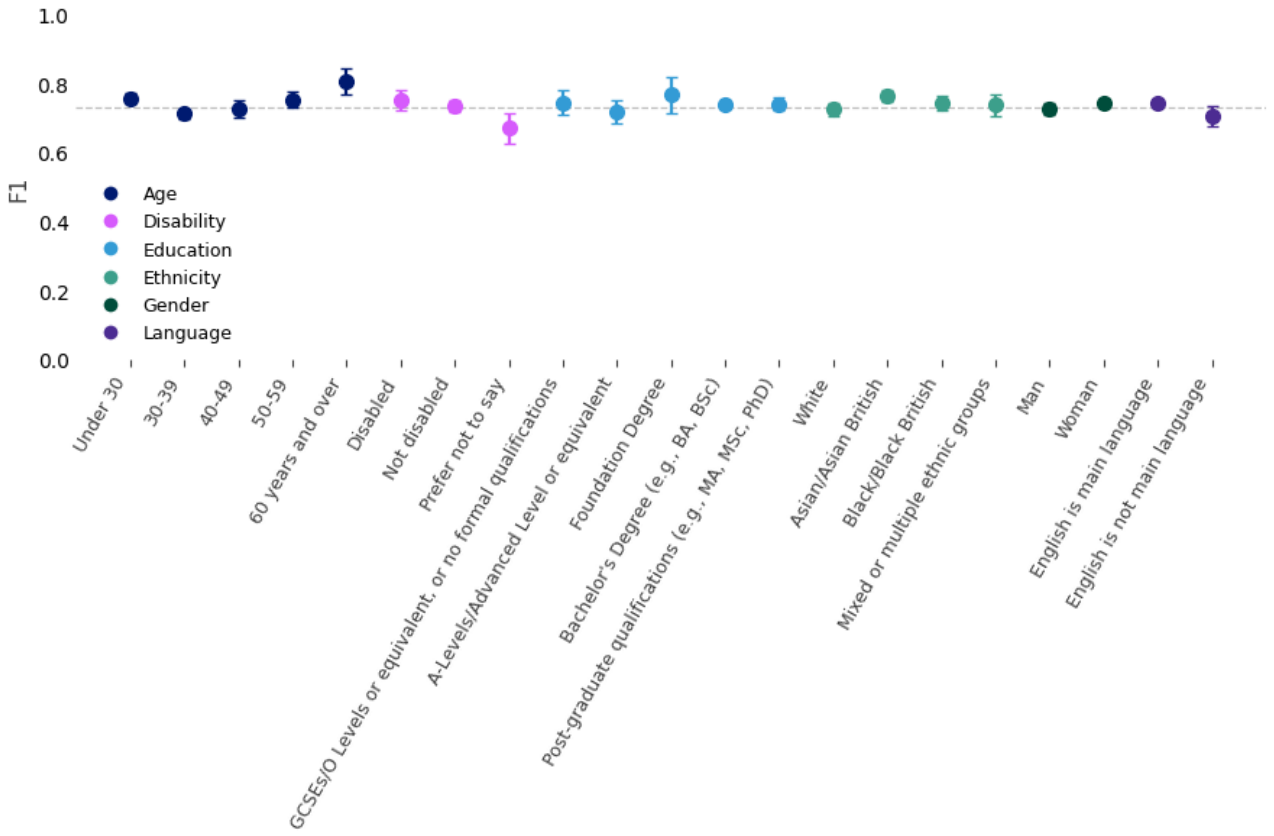
where $\mathbf{X}_i$ is a vector of categorical covariates capturing respondent age category, disability status, education level, ethnicity, gender, and language (see Figure 9 for specific categories), and $\delta_{q(i)}$ denotes question fixed effects.

As communicated in Section 1.3.3, our principal concern here is that (1) respondents in certain demographic groups may systematically use (some) different language and (2) LLMs may perform worse at interpreting this language relative to human analysts.

8.3 Demographic bias results

Figure 9 presents the average F_1 results by demographic, illustrating similar results within each group (minimum per category=36; bootstrap confidence intervals shown).

Figure 9 Theme mapping (mean F_1) by demographic characteristic



We did not find evidence of systematic bias in the regression results reported in Annex 6. The adjusted R^2 was 0.001, with an overall F-statistic of 0.46, indicating that demographic characteristics explained negligible variation in the overall accuracy of theme mapping. The regression coefficients for each demographic characteristic (see Annex 6) generally show negligible practical differences in F_1 accuracy. Although the negative coefficient for language, assuming those with English as an additional language have 'poorer' written English, is consistent with the concern that responses written in poor English are associated with degradation in accuracy, these results were not practically or statistically significant. Some statistically significant differences were observed for ethnicity; compared to White ethnic groups, average accuracy was statistically significantly higher for Asian/Asian British and Mixed/multiple ethnic backgrounds. However, the practical significance was low, with a difference of less than 3 percentage points and a standard deviation (SD) effect size < 0.12. Taken at face value, these results suggest that minority ethnic groups are associated with slightly higher F_1 scores compared to White respondents on average, a direction which is contrary to typical algorithmic bias patterns. There were no statistically or practically significant results by age, disability, education, gender, or language.

We also completed sensitivity analysis by examining all 3 questions separately and by three different accuracy-based metrics: F_1, recall, and precision. The results from nine regression models (not shown) did not undermine the above conclusions; results varied by question and outcome, revealing no consistent pattern. For example, the coefficient for Asian/Asian British and Mixed/multiple ethnic groups was not statistically significant for most of the models and the direction of the coefficients varied across models. The conclusions were insensitive to alternative models tested, including logistic regression where F_1 was converted into a binary variable.

We acknowledge that this analysis is not exhaustive and faces limitations related to the data source itself. Firstly, our data collection was not well represented amongst those with no formal qualifications (7 respondents), minority gender categories (7), and those aged 70 and over (11) so we could not analyse differences for these groups. We also collected detailed disability types but, again, there were too few respondents to draw useful insights. Secondly, the respondents were all recruited via Prolific,[57] which suggests a minimum level of English and technical proficiency, meaning respondents are not necessarily representative of the public with the same observable demographics. Thirdly, although demographic information was hidden from the human analysts that created the reference data, it is feasible to infer demographics from language (e.g., use of vernacular more commonly associated with certain age groups) and thus the ground truth itself may contain bias as the true themes themselves could be influenced by human analysts' (un)conscious bias. In addition, this analysis does not assess potential bias in the theme generation stage and bias is a concern at this stage too (Ashwin et al., 2025).[58]

[57] https://www.prolific.com

[58] Ashwin, J., Chhabra, A., & Rao, V. (2025). Using Large Language Models for Qualitative Analysis can Introduce Serious Bias. *Sociological Methods & Research*, *0*(0). https://doi.org/10.1177/00491241251338246

9. User feedback

9.1 Summary

Throughout the live pilot, the team collected feedback from users via surveys (n=15), structured feedback, and face-to-face meetings. Users included policy officials, statisticians, social researchers, and customer insight officers. Two-thirds of users were experienced in thematic analysis. Feedback was positive overall, with all respondents either "likely" or "very likely" to recommend the CAT.

"The CAT was invaluable and enabled us to dedicate our efforts to more focused data (analysis)." (CAT user)

"Thanks for saving hours of manual coding!" (CAT user)

"It saved time and was useful." (CAT user)

"I found it (theme review) fairly straightforward." (CAT user)

9.2 Human-in-the-loop process feedback

Users were generally positive about the human-in-the-loop theme review. The majority (60%; n=15) reported that it was easy, with the remainder reporting somewhat easy.

Users found in-person training helpful when preparing for the theme review, where they were shown how social researchers usually code qualitative data, enabling the review to be completed in a structured manner. They took a (self-reported) average of 160 minutes (min=45 mins; max=5 hours) to complete the theme review per 100 responses reviewed. The length of responses and number of themes meant this process varied substantially, as reflected in one user's comment: "*themes were quite straightforward to review, although some responses required longer to analyse*". Users regularly wanted to discuss with a colleague whether a theme should be changed, dropped, or added, which could sometimes take significant time.

Explainability was also valued by users; for example, users liked being able to trace the consolidated themes back to the initial themes generated from multiple ‘analyst’ passes over the same dataset (see Section 3.2).

9.3 User challenges feedback

Users were asked to report any themes that they found difficult to review and any other challenges.

- **Theme overlap**: There were instances where there was some conceptual overlap between CAT-generated themes, meaning users had to decide how to best rationalise e.g., drop one theme or combine them.

- **Theme hierarchy:** Some reviewers reported struggling to decide on theme hierarchy, particularly distinguishing between overarching themes and sub-themes, e.g., deciding whether to have a general theme on accessible travel versus several mode-specific accessible travel themes.

- **Incorrect interpretation:** Some examples were identified where the CAT misinterpreted the sentiment and/or reported the presence of unsubstantiated themes, leading to inconsistencies in the mapping step. One user suggested it would be helpful to communicate any limitations in the AI to help them be particularly vigilant about spotting these cases.

- **Clarity in guidance:** Some users requested clearer guidance on golden insights, particularly the conceptual distinction from main themes. Some users also wanted more clarity on whether they were *required* to publish the bootstrap confidence intervals provided for the CAT theme mapping results used to illustrate uncertainty.

10. Lessons learned

10.1 Summary

In this section, we summarise key lessons learned regarding benefits, methodology, evaluation design, human-in-the-loop design, and opportunities for improvements.

10.2 Benefits

- **Cost savings:** As the CAT replaces the majority of the manual work involved in thematic analysis, we estimate that it saves around 50-70% of the entire cost of responding to a medium-sized consultation. The remaining fixed costs include human review, survey design, project management, synthesis, and report writing. On the four consultations completed to date[59] (which varied significantly in size), the CAT has analysed 200,000 responses and over 8 million words, saving an estimated £0.5 million compared to a scenario where all responses for all consultations were rigorously analysed by humans. The DfT conducts roughly 55 consultations annually; assuming that the cost of a medium-sized consultation is £80-100k and applying the above estimated saving percentages then this could yield savings in the range of £1.5-4 million per year.[60] Here, we define medium-sized as around 1 million words and 15,000 responses and we use rough estimates of survey costs, feedback from social researchers who have completed qualitative coding, and cost and time estimates of using the CAT.

- **Time savings:** The CAT saves significant full-time equivalent (FTE) hours. The human-in-the-loop theme review takes just 1-5 hours per question; even with this investment, the CAT has roughly saved 15,000 hours of work to date compared to a scenario where all responses for all consultations are rigorously analysed by humans manually. This represents a significant 'opportunity cost' saving, so time can be invested into additional activities that support keeping the UK on the move.

[59] As well as the four projects (consultations [x1], Call for Evidence [x1], and Call for Ideas [x2]) described in Annex 2, the CAT was used on the 'A railway fit for Britain's future' consultation. However, officials simply used the CAT-generated themes to cross-reference themes that they had already identified through the manual analysis of consultation responses.

[60] This includes consultations and Calls for evidence (GOV.UK). If we use the same approach that others have used to estimate the average savings per government consultation from AI, our central figure is £2.2 million (i.e., within the stated range): https://ai.gov.uk/knowledge-hub/tools/consult/

- **Responsiveness:** As the CAT significantly speeds up analysis, the DfT can extract the insights required to inform evidence-based policy and delivery more quickly.

- **All sizes of consultation benefit:** The size of the time and cost savings benefits increases for larger consultations. However, even on small consultations (< 50 responses) where users were required to read all responses (rather than a random sample) during the human-in-the-loop review (see Annex 3), users of the CAT reported time savings and found the consistent structure of the analysis helpful.

- **Completeness:** Some government consultations do not analyse every single response for very large consultations. The CAT does not face resource constraints – every response is always processed for both the theme generation step and the theme mapping step.

- **Quality:** The performance evaluation results demonstrate that the CAT delivers high-quality analysis compared to human-analysed datasets. It is also likely that the CAT leads to lower error as the risk of human fatigue-related error is reduced, something that members of the public anticipated (DfT, 2023).[61]

- **Human-centred design:** DfT (2023) research on public attitudes towards using AI for consultations emphasised the importance of human oversight. Responding to this, our evaluation approach demonstrates that the CAT aligns with human analysis and our human-in-the-loop design helps mitigate the risk of errors.

10.3 Methodology learnings

- **A pure LLM method worked best:** Using a 'pure' LLM pipeline approach for theme generation and mapping achieved similar or better results than alternative methodologies tested that blended LLMs with more traditional machine learning methodologies, such as clustering (see Annex 4).

- **Conveying mapping uncertainty:** We expected to find a relationship between models' token-level log probabilities (logprobs) and their likelihood of misclassifying theme labels relative to a human, offering a potential way to communicate uncertainty to users. Instead, we observed minimal variation in logprobs across classification tasks: the model appeared approximately certain on all Boolean mapping classifications that we tested, even when it was deemed incorrect to human reviewers. This outcome reflects that logprobs capture the model's confidence in generating a particular (constrained) token (e.g., 'True'), conditioned on the prompt, rather than its epistemic certainty of that classification decision. We chose to present bootstrap confidence intervals to communicate aggregate uncertainty in the *estimated* theme prevalence rates instead as the CAT outputs aggregate statistics alongside the raw analysed data.

- **Ensembles of LLMs boosted performance:** Using a mixture of LLMs, where each data pass randomly changed the position of responses, improved the theme

[61] Department for Transport. (2023). Public attitudes to the use of AI in DFT consultations and correspondence. https://assets.publishing.service.gov.uk/media/654e6f078a2ed4000d720d12/using-ai-in-consultations-and-correspondence.pdf

generation performance. Using a majority vote LLM ensemble method improved theme mapping performance, particularly for theme rank evaluation metrics.

- **Judgements on theme granularity were necessary:** Designing the CAT required judgements on how thematic analysis 'should' be done, including how a theme should be defined and at what level of aggregation. Engaging social researchers for advice was critical here. However, this led to challenges where the granularity of how the CAT analysed themes differed from some human ground truth datasets; for example, we tested the CAT on a number of historical consultations which identified extremely general themes compared to the CAT, such as "Benefits (general)" versus more specific benefits that the CAT drew out as themes. One feature we intend to explore in the future is a hierarchical theme framework for main themes, comprising sub-themes, themes, and overarching themes.

10.4 Evaluation design learnings

- **Ground truth is slippery in qualitative analysis:** Language is subjective, experts legitimately disagree, and coding errors or variability creep into qualitative analysis. As a result, comparisons to human-analysed datasets are inherently contestable, which makes evaluation challenging. For example, variability in the CAT's accuracy results across datasets is partly a function of the variability in the quality and heterogeneity of the human-analysed reference datasets themselves. This renders a perfect evaluation accuracy score of 1 as practically unachievable for the CAT in many cases.

- **Having an automated evaluation pipeline aided development:** An automated evaluation pipeline, with a diverse ground truth dataset, provided a pragmatic, rapid, and robust approach to test methodology changes and new features. It also proved useful for regression testing. The development of the evaluated LLM-as-a-Judge was crucial here to avoid human intervention every time we wanted to evaluate changes to the theme generation evaluation pipeline.

- **Developing effective LLM-as-a-Judge evaluators requires human-AI iteration:** LLM-as-a-Judge evaluators need to be designed and validated in an iterative loop to maximise their usefulness. For this, we used a varied human-analysed ground truth dataset and iterated on the criteria by reviewing judge results on this dataset. Binary classification performed best.

- **Controlling randomness during AI evaluation is critical:** Because LLMs produce stochastic outputs, fixing random seeds (data sampling) and sampling hyperparameters (temperature, Top-P) across multiple runs in our automated evaluation pipelines enabled more reliable isolation of performance impacts from feature modifications, prompt engineering, model upgrades, and tuning.

- **LLM sensitivity:** As many others have identified (cf. Shankar et al., 2024),[62] the LLM results were sensitive to minor changes in the prompts. Prompts needed to be re-engineered and the pipeline re-evaluated when model and version upgrades

[62] Ibid

were introduced. This underscores the importance of off-the-shelf AI tools being thoroughly evaluated for the task at hand.

10.5 Human-in-the-loop design learnings

- **Use Bayesian updating to inform the number of responses to review:** A key challenge was deciding how many responses should be reviewed when conducting the human-in-the-loop theme review. We developed an explainable, assumption-based model (see Annex 3) to estimate the number of responses that should be reviewed to be approximately 95% confident that no important themes were missed. This model provides a prior expectation on how many responses to review. Users can then update this prior expectation based on the qualitative feedback loop from the act of reviewing the data itself. This helps users plan, invest scarce resource most pragmatically, and make informed decisions.

- **Users value training:** Although users generally found the process easy, as shown in the user feedback (see Section 9), improved guidance, more hands-on training and support would be beneficial.

10.6 Opportunities for enhancements

- **Use LLM-as-a-Judge for self-corrections:** Validation steps within the pipeline could be used to diagnose and self-correct issues. For example, LLM-as-a-Judge could assess whether the theme consolidation criteria are met and automatically re-run the process if not.

- **Diagnosing weaknesses:** More evaluation work could be undertaken to understand the determinants of relatively low performance for certain datasets and themes, such as investigating associations between performance and: types of themes, typology of questions, and lengths of questions.

- **Mapping sources of uncertainty at inference time:** Future work could explore confidence-based quality gates to help prioritise targeted human review. For example, we could flag themes where the model exhibits low confidence during theme generation steps. For the mapping step, IRR scores can be automatically derived from LLM ensemble results to identify any themes or responses with high variance, possibly indicating responses requiring expert interpretation.

- **Feature enhancements:** Users would benefit from improvements to user experience design, hierarchical theme extraction, and improvements to user guides and training.

- **Additional safety features:** Ongoing safety work includes adversarial red teaming exercises to assess and mitigate prompt injection risks, including scenarios in which malicious survey responses attempt to influence the LLM's analysis.

11. Annex

Annex 1. Data used for the (blind) automated evaluation overview

All responses (9,100) to all questions (11) below were used for the blind evaluation design. Human-vs-human mapping IRR scores were calculated for these three overall datasets.

Dataset ID	Background stats	Dataset curation information	Human-vs-human inter-rater reliability (IRR)
INTS	**# questions:** 3 **# responses per question (mean):** 200 **Response word count (mean):** 36 **# of 'true' themes per question (mean):** 11	For one question (INTS_Q48) taken from the INTS Call for Ideas, a random sample of responses (200) was extracted from a raw dataset and coded by a researcher. Researchers coded 50 responses to generate a codebook, which they then used to independently code the next 50 responses. When new codes were identified, these were discussed and added to the codebook. Cohen's kappa was calculated on a sample of 50 separately coded responses (0.77), which showed good IRR, suggesting that it was suitable for one researcher to continue coding the remaining dataset. Due to practical constraints, and given the IRR score previously achieved, a faster approach was used to analyse the other two questions (INTS_Q31 and INTS_Q38). Each researcher developed the codebook and analysed all the responses to one of the remaining questions. The researchers then reviewed 10% of the other's coded responses. The IRR scores for the remaining two questions were 0.99. Codes (sub-themes) were then aggregated into themes. Inductive framework analysis was used, whereby codes and themes were not predetermined but emerged from the data.	**Q48 overall proportion agreement**=0.97 **Q48 macro proportion agreement**=0.79 **Q48 Cohen's kappa**=0.77 **IRR sample**=50 **Number of coders**=2
ATI_A	**# questions:** 5 **# responses per question (mean):** 977 **Response word count (mean):** 30 **# of 'true' themes per question (mean):** 12	This dataset was derived by a mixture of work by The Alan Turing Institute (ATI) researchers and crowd-workers procured via Prolific. 1,000 crowd-sourced workers answered five free-text questions to create the raw data. For each question, themes were then identified by ATI and crowd-workers and then all responses were annotated by three trained crowd-workers. 11% of the original dataset was dropped where workers failed attention tests. For our mapping analysis, true theme mappings were defined as majority vote (>= 2 annotators assigned the same theme). The kappa results shown communicate the average pairwise Cohen's kappa of each annotator with this majority vote dataset.	**Average overall proportion agreement**=0.96 **Average macro proportion agreement**=0.82 **Average Cohen's kappa**=0.67 **IRR sample**=4,886 **Number of coders**=3
ATI_B	**# questions:** 3 **# responses per question (mean):** 1,204 **Response word count (mean):** 40 **# of 'true' themes per question (mean):** 22	1,200 crowd-workers were used. Demographics info was captured on ethnicity, language, sex, gender, age, disability and education. Theme identification and annotation followed the same procedure as the ATI_A data (for rapid model development). As with ATI_A, majority vote was used to define true theme mappings for our analysis.	**Average overall proportion agreement**=0.96 **Average macro proportion agreement**=0.80 **Average Cohen's kappa**=0.71 **IRR sample**=3,612 **Number of coders**=3

Table notes:
Overall proportion agreement is measured as the overall agreement of all labels (every classification label instance is equally weighted). Macro proportion agreement weights theme presence and theme absence equally. The difference between these metrics demonstrate that there is more agreement amongst instances of theme absence than instances of theme presence.
Rare themes are included in (main) themes here – we do not distinguish between golden insights and main themes for these datasets.

Annex 2. Data used for the (non-blind) live pilot evaluation overview

All 32 questions (total responses=198,000; 480 'true' themes; 749 'true' golden insights) summarised below were used in the live pilot theme evaluation of theme generation across four consultations and Call for Evidence surveys. Of these, a random sample from 7 questions (DVSA and Ukraine) were used for the theme mapping evaluation (total responses=720).

ID	# of questions	# of responses (mean per question)	Response word count (mean)	# of true main themes (mean)	# of true golden insights (mean)	Human-vs-CAT inter-rater reliability (IRR)	Dataset information
DVSA	6	25,600	28	16	64	Proportion agreement IRR=0.99; Cohen kappa=0.91; IRR sample=600	Respondents submitted their answers to questions on a consultation to improve car driving test booking rules.[63] The team (policy and social researchers) in DVSA used the CAT, completing the human-in-the-loop theme review. A social researcher reviewed a random sample of 100 responses per question to review the CAT theme mapping, changing any label they disagreed with (main themes only).
INTS	19	2,300	89	14	15	-	Respondents submitted their answers to questions on the Integrated National Transport Strategy (INTS) Call for Ideas.[64] The team (policy officials and social researchers) used the CAT, completing the human-in-the-loop theme review. No IRR was calculated (note that IRR scores were calculated for CAT-vs-human analysis for 3 INTS questions as part of the automated [blind] evaluation).
NZP	6	34	260	17	10	-	Respondents submitted their answers to questions on the Net zero ports: challenges and opportunities call for evidence[65] The team (policy officials and social researchers) used the CAT, completing the human-in-the-loop theme review. No IRR was formally calculated but all responses and mapping analysis was reviewed by the team and positive qualitative feedback on the accuracy was provided.
UKR	1	377	71	13	24	Proportion agreement IRR=0.98; Cohen kappa=0.92; IRR sample=120	Respondents submitted their answer to a transport-related survey from residents in Kharkiv. Responses were translated from Ukrainian into English by translators at the Foreign, Commonwealth and Development Office. DfT policy officials used the CAT, completing the human-in-the-loop theme review. IRRs were calculated for human-vs-CAT (as shown) and human-vs-human (social researchers) to contextualise IRR findings (see Annex 5).

[63] https://www.gov.uk/government/consultations/improving-car-driving-test-booking-rules

[64] https://www.gov.uk/government/calls-for-evidence/integrated-national-transport-strategy-a-call-for-ideas/integrated-national-transport-strategy-a-call-for-ideas

[65] https://www.gov.uk/government/calls-for-evidence/net-zero-ports-challenges-and-opportunities

Annex 3. Human-in-the-loop theme review sample analysis

Summary

For the CAT v1.0 live pilots, we developed an explainable, assumption-based model to estimate the minimum number of responses that should be reviewed to be approximately 95% confident that no main themes were missed.

Users provided information about a dataset, including number of responses and their initial expectation of the number of 'true' main themes they expect to be in the dataset (see illustrative example below). Previous consultations and the CAT theme generation itself provide information to support this expectation. The model provides an estimated initial number of responses that should be reviewed and users can then *update* this prior expectation based on the qualitative feedback loop from the act of reviewing the data itself; for example, if after reviewing the recommended number of responses they identified additional themes that the CAT missed then they should continue reviewing responses until they deem theme saturation has been achieved. This approach combines quantitative insights from the simulation modelling and the concept of theme "saturation" (see Section 1.3.1), which is a subjective criterion regularly used in qualitative thematic analysis.[66]

This is essentially a Bayesian updating approach, where the model provides a prior which is then updated based on the qualitative experience of the theme review itself. This helps users plan, balance risk and pragmatism, and make informed decisions.

Simulation and regression method overview

We created a large (n=200,000) simulated dataset where *each example* (row) in the dataset was constructed using the following steps:

- **Step 1 (define consultation dataset characteristics):** We randomly define the characteristics of the hypothetical consultation, including: number of 'true' themes, number of responses, distribution of number of themes *within* responses, and distribution of number of themes *across* responses.
- **Step 2 (define HITL process):** We define how many responses were reviewed in a random sample of responses and what percentage of all the true themes were identified following this simulated human-in-the-loop theme review empirically.
- **Step 3 (calculate perfect recall):** Where >= 95% of true themes were identified from the above process, we defined this as perfect recall.

This yielded a rich and varied dataset where the characteristics of each consultation and human review were realistic. We used probabilistic assumptions based on historical consultations and literature to achieve this.

We then developed an explainable model (logistic regression); perfect recall was regressed on key features (see illustrative example below) constructed during the simulation. The resultant

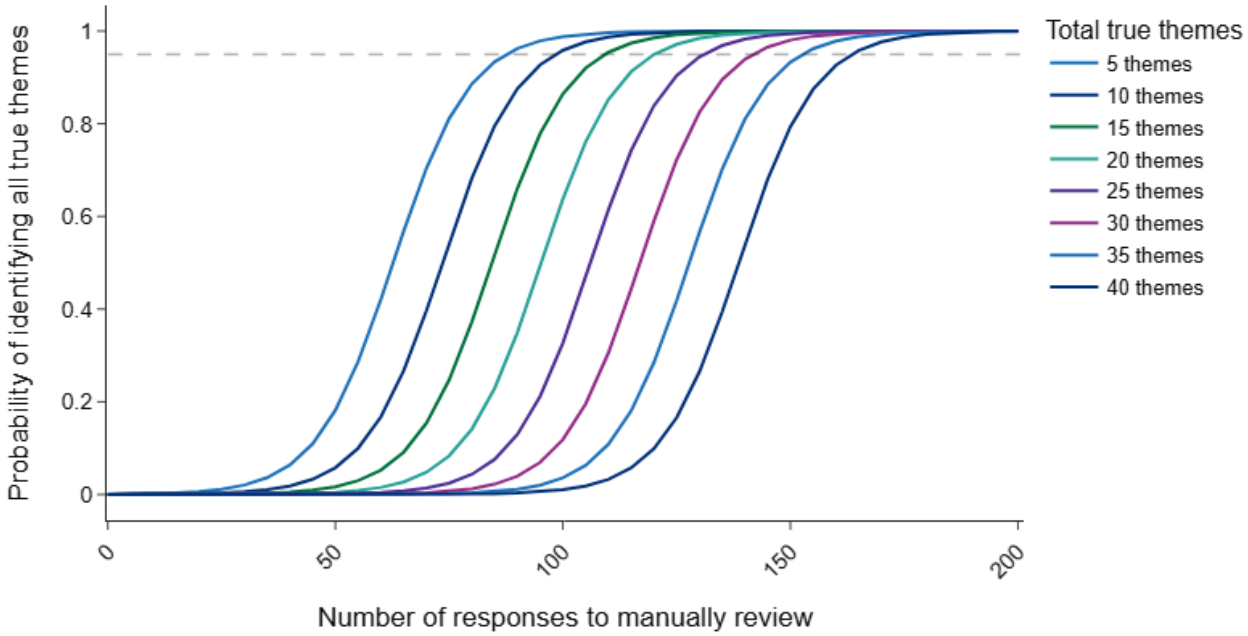
[66] Saturation in qualitative research: exploring its conceptualization and operationalization

coefficients were used in a flexible calculator so users could estimate how many responses *should* be initially reviewed to be approximately 95% confident that perfect recall is achieved.

Illustrative example

Figure 10 is an illustrative visualisation from the calculator, where the following assumptions are used as model inputs: 5,000 total responses, the assumed minimum theme prevalence rate of a single theme is 5%,[67] and the maximum number of themes that can exist in a single response is five. Under these assumptions, the model indicates that if 15 true themes are present, a user should initially review a random sample of 120 responses to be 95% confident that all themes 'show up' at least once in the sample. Human reviewers then use their subjective judgment based on qualitatively reviewing the data to decide whether theme saturation has been met or not.

Figure 10 Human-in-the-loop theme review sample calculator example



Annex 4. Additional methodologies explored

Summary

During early stages of development, we developed and evaluated alternative methodologies before deciding on the CAT v1.0 approach. Our aim was to develop a methodology that used the power of LLMs, so all methodologies explored relied heavily on LLMs. The CAT v1.0 methodology is a pure LLM-approach where a pipeline of LLMs is used for all aspects of theme generation and theme classification (mapping). However, we also explored a range of LLM-blended approaches which used a combination of LLMs and clustering. Although the results below illustrate that the alternative methods performed well, we concluded that our 'pure' LLM-based approach was preferable early on during development. It delivered more consistent recall, provided more flexibility for feature enhancements to be added, and offered

[67] Savanta have suggested that main themes (categories) should be represented in a minimum of 5% of responses. The results are sensitive to this assumption.

greater scope for performance gains given the pace at which increasingly capable LLM models are released by AI labs.

Alternative LLM-blended methods explored

Responses were pre-processed into a set of low-level insights, breaking up each response into its component arguments and points using an LLM. These low-level insights were then converted into text embeddings to allow for numerical clustering algorithms. After clustering, an LLM summarised the central responses of each cluster to identify the theme. We tested the following approaches.

- **K-means:** Using the low-level insight embeddings, an LLM identifies a number of clusters and initial starting centroids. The K-means algorithm then searches for a final set of cluster locations which minimises the within-cluster sum-of-squares metric.

- **Hierarchical clustering:** Using the principal components of the low-level insight embeddings, the Agglomerative clustering algorithm was used to build up clusters from the bottom in a deterministic way. To automatically choose a cutoff point (number of themes), 30 options were selected based on their linkage distance percentile, and the silhouette score was used to determine optimum distances from values.

- **HDBSCAN:** Using the principal components of the low-level insight embeddings, the HDBSCAN algorithm was used to identify clusters based on areas of high density which have a minimum size based on a percentage of the dataset.

- **LDA:** Using the low-level insight text, Latent Dirichlet Allocation (LDA) creates topics based on n-gram occurrences within the low-level insights. A wide range of topic numbers were tested (based on the size of the dataset), with the coherence and stability measures used to automatically determine the number of themes.

Figure 11 Results for alternative LLM-blended methodologies tested (theme generation results)

Method	Precision	Recall	F_1
K-means	0.77	0.81	0.77
Hierarchical	0.89	0.58	0.68
HDBSCAN	0.77	0.64	0.67
LDA	0.81	0.62	0.67

Annex 5. Inter-rater reliability scores

Figure 12 summarises the theme mapping inter-rater reliability (IRR) statistics for questions where Human-vs-Human IRRs were created alongside CAT-vs-Human IRRs using the metrics of kappa and proportion agreement described in Section 1.3.1. Comparable IRR statistics were only available for the below questions in our evaluation due to the time-consuming nature of developing IRR statistics. As can be seen from the below, all CAT-vs-Human IRR scores had an overall agreement rate of > 88% and the CAT-vs-Human IRR scores were generally similar or higher than the Human-vs-Human IRRs, except INTS_Q48.

Note that the way the IRRs are derived below for ATI_A and ATI_B are *not comparable* to the data as described in Annex 1 and analysed in Section 7 where we instead use a 'majority vote' rule from multiple human analysts to define the human-analysed reference data.

Figure 12 IRR mapping results for human-vs-human and CAT-vs-human

Dataset	Descriptive stats	Human-vs-Human IRR			CAT-vs-Human IRR		
	Number of responses	Kappa	Overall proportion agreement	Macro proportion agreement	Kappa	Overall proportion agreement	Macro proportion agreement
Overall	**8,598**	**0.52**	**0.92**	**0.70**	**0.56**	**0.92**	**0.72**
ATI_A_Q1	995	0.49	0.94	0.76	0.53	0.94	0.78
ATI_A_Q2	992	0.51	0.94	0.73	0.54	0.94	0.73
ATI_A_Q3	950	0.40	0.92	0.65	0.43	0.92	0.67
ATI_A_Q4	959	0.46	0.89	0.68	0.50	0.89	0.69
ATI_A_Q5	990	0.34	0.91	0.63	0.37	0.90	0.64
ATI_B_Q1	1,204	0.54	0.93	0.70	0.56	0.92	0.70
ATI_B_Q2	1,204	0.48	0.94	0.68	0.53	0.94	0.70
ATI_B_Q3	1,204	0.46	0.91	0.65	0.50	0.91	0.66
UKR_Q1	50	0.75	0.91	0.79	0.92	0.98	0.94
INTS_Q48	50	0.77	0.97	0.79	0.60	0.89	0.69

Table notes:
Overall proportion agreement is measured as the overall agreement of all labels (each classification of each theme absence and theme presence classification is equally weighted).
Macro proportion agreement weights theme presence and theme absence equally.
For datasets ATI_A and ATI_B, three human annotators analysed each question which required adjusted metrics. The proportion agreement rate for these datasets takes the average of pairwise comparisons. For the Human-vs-Human IRR, each of the three annotator's analysis is compared to one another and then averaged across all possible (3 comparisons) pairwise comparisons. For the CAT-vs-Human, we take the average pairwise comparison between the CAT and each annotator (3 comparisons). Fleiss's kappa is used for our kappa metric for these datasets. Results were insensitive to similar, alternative metrics such as average pairwise Cohen's kappa.

Annex 6. Demographic bias regression results

Figure 13 Theme mapping bias evaluation regression results by characteristic

Category	Characteristic	N (Per Question)	Coefficient	SD Effect Size	Standard Error
Intercept	-	-	0.70	-	-
Age	60 years and over	57	0.02	0.08	0.02
	50-59	120	0.01	0.03	0.02
	40-49	196	0.01	0.04	0.01
	30-39	412	0.00	0.01	0.01
	Under 30	417	*Reference*	-	-
Disability	Prefer not to say	43	0.00	-0.02	0.02
	Disabled	226	0.01	0.05	0.01
	Not disabled	935	*Reference*	-	-
Education	Post-graduate qualifications (e.g., MA, MSc, PhD)	336	-0.01	-0.03	0.01
	Bachelor's Degree (e.g., BA, BSc)	524	0.01	0.05	0.01
	Foundation Degree	46	0.03	0.11	0.02
	GCSEs/O Levels or equivalent, or no formal qualifications	76	0.01	0.05	0.02
	A-Levels/Advanced Level or equivalent	208	*Reference*	-	-
Ethnicity	Black/Black British	270	0.02	0.06	0.01
	Asian/Asian British	307	0.03*	0.11	0.01
	Mixed or multiple ethnic groups	202	0.03*	0.12	0.01
	White	387	*Reference*	-	-
Gender	Man	576	-0.01	-0.03	0.01
	Woman	614	*Reference*	-	-
Language	English is not main language	129	-0.01	-0.03	0.02
	English is main language	1,071	*Reference*	-	-

Table notes:
The dependent variable is response-level F1 (n=3,612 across all three pooled questions). The dataset is ATI_B (dataset details described in Annex 1).
The table shows the specification with gender included (a regression with sex included instead yielded extremely similar results).
*Coefficient significance thresholds (using robust standard errors): * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$*