Incubator for AI

Consult Evaluation: Scottish Government's Non-surgical cosmetic procedures consultation

20/03/25 | Lucy Makinson, James Lowe, Alex Moore, Michael Pryse-Davies, Harry Banton, Katie French, Charlotte Ryall, Nina Menezes, Rachael Robinson, Jude Webb, Rishav Punatar

Executive summary

Consult is a tool for analysing public consultations, which uses AI to generate themes for each question and map responses to those themes. It alleviates the burden of running consultations by facilitating quicker analysis for a fraction of the cost.

This report outlines findings from an evaluation of Consult using a Scottish Government consultation on the regulation of non-surgical cosmetic procedures.

Headlines:

- ➔ **Consult was generally good at mapping responses to themes and produced near-identical rankings of themes compared to expert reviewers.**

  The overall performance (F1) score was 0.76 out of 1 (generally considered 'good') and reviewers made no changes to the Consult themes for 60% of responses.

  Differences between Consult and reviewer-identified themes had negligible impact on the overall theme rankings – the key driver of policy recommendations.

  Changes to themes were heavily concentrated on a small number of specific themes. This should make it quicker to detect and correct difficult themes early.
- ➔ **Reviewing Consult themes is quick, with a median time of 23s per response. By reducing analysis time, Consult freed up time to focus on the implications.**
- ➔ **Reviewers were impressed with the initial themes generated and felt that Consult helped reduce reviewer bias. However, they still wanted an opportunity to influence the themes, and the process for this was cognitively demanding and time intensive.**
- ➔ **Consult struggled to identify missing themes. The theme generation step may require more iteration, and working with the theme longlist is likely to help.**
- ➔ **Reviewers were keen to move to a world where only a share of responses need reviewing, but needed more confidence in the mapping before doing that.**

Contents

Introduction

Background

What is the incubator for Artificial Intelligence (i.AI)?

i.AI rapidly design, test and deliver AI products for government. We take pride in being a highly technical team, composed of experts with deep knowledge and experience from both the public and private sector. This allows us to tackle complex challenges and innovate continuously towards our overall mission: to harness the opportunity of AI for public good.

What is Consult?

Each year the government runs around 600 public consultations, with some receiving more than 100,000 responses. Analysing these consultations takes hundreds of thousands of hours of civil service time, or are contracted out at substantial cost. The time-intensive process also delays the findings and, ultimately, the development of policy.

Consult, an AI-powered tool developed by i.AI, aims to alleviate this burden by facilitating quicker analysis, of comparable quality, for a fraction of the cost. It does this using a two stage process:

1. 'Theme Generation': Identify common themes in the responses to generate a list of themes for each question.
2. 'Theme Mapping': Classify the full set of responses using one or more of these generated themes.

Why are we running this evaluation?

Consult has the potential to save substantial time and money for Government, but it is critical that this is not at the expense of high quality analysis that captures the public's views. We are therefore carrying out several evaluations to ensure that Consult is suitable for wider use, and continually improving its delivery. We have conducted several retrospective evaluations using historic consultation data. This evaluation builds on our previous evaluations in several ways:

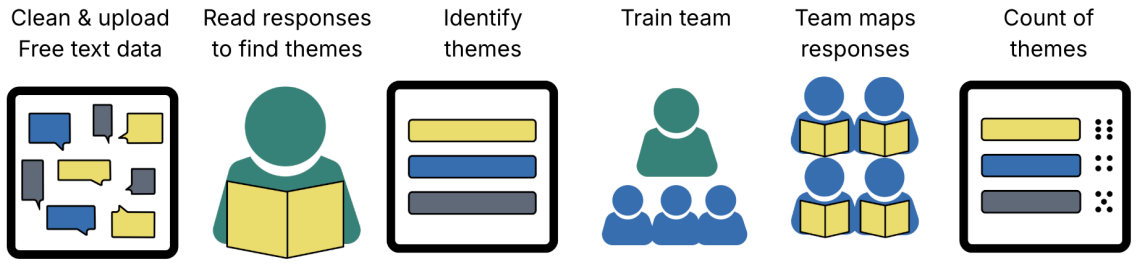
1. This is the first evaluation we have done on a live consultation, so allows us to better understand how Consult performs in practice.

2. It is the first evaluation we have done with the current version of Consult. This includes a new approach to reviewing the initial theme list (generation stage), and a new interface for reviewing mapped themes (mapping stage).
3. It addresses a new topic. In order to be confident in Consult's use for consultation analysis in general, we need to evaluate it across a range of different topics.

How does Consult work?

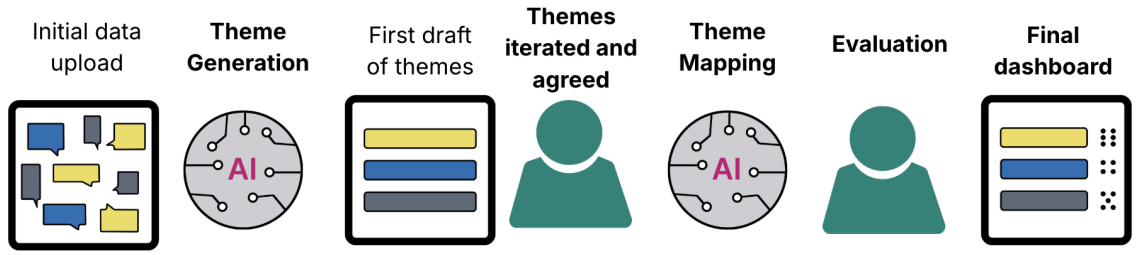
How teams currently analyse consultation responses

Conventional analysis of public consultations is a multi-stage process. Although the approach varies by consultation (for example, based on their size) it generally involves an initial review of a subset of responses to identify common themes. Once these are identified, a team is trained on the theme framework (to ensure they have a common understanding of each theme) before reading through all of the responses and mapping each response to the list of possible themes. At the end, a summary shows how frequently each theme appeared.



How Consult analyses responses with a human in the loop

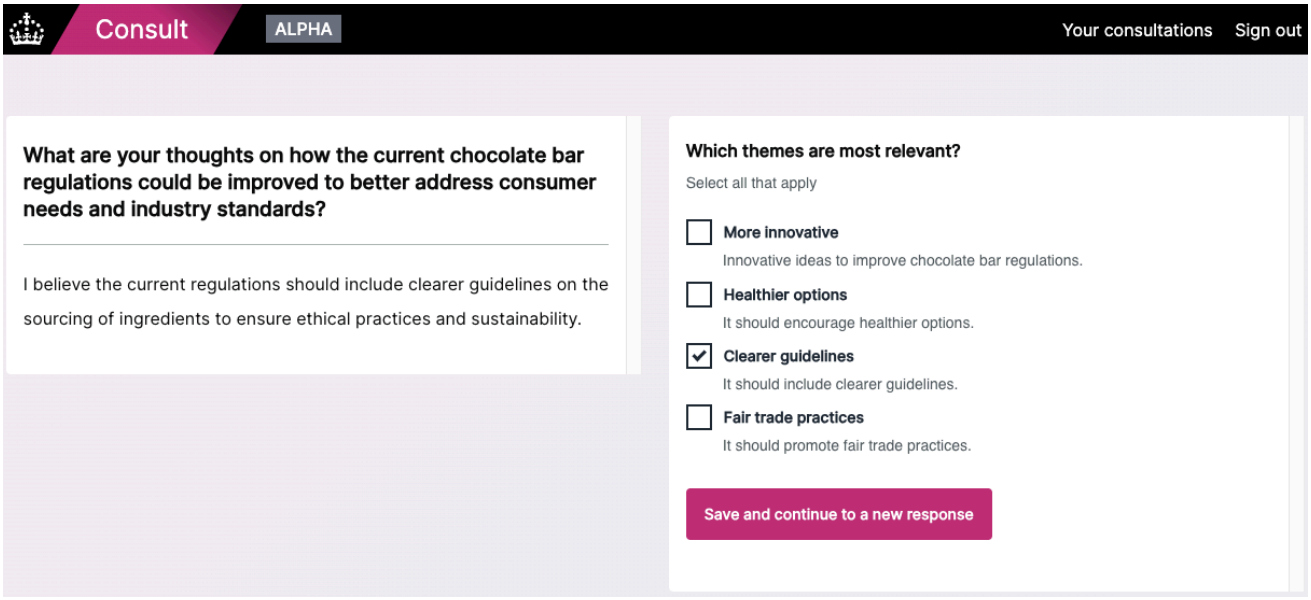
Consult broadly follows two steps: theme generation and theme mapping. First, it uses a topic modelling approach to identify common themes for each question (using the whole dataset, rather than a sample). It then maps each response to the available themes using LLMs. We are developing a dashboard to summarise the findings, allowing policymakers to see how often each theme appeared, as well as more nuanced metrics such as the overlap between themes.



This two stage process allows us to include a 'human in the loop'[1] at each stage. In the first stage (theme generation), expert reviewers working on the policy area can review and approve the themes before they are used in the mapping process. This theme sign-off process is a new development that we were keen to test. *For this evaluation, this stage involved four expert reviewers from the Scottish Government.*

For the second stage (theme mapping), reviewers verified (and, if necessary, amended) the theme mappings that Consult provided for each response. Using the Consult application, reviewers were presented with an individual free text response to an individual question and the themes that Consult had mapped to it. If they deemed it necessary, reviewers made changes to the themes which Consult had mapped to the response using the tickboxes. Once they were content with the set of themes mapped to the response, they saved the mapping and continued to the next response. You can see a hypothetical example below. *For this evaluation, this stage involved six expert reviewers, who split the responses between them.*

Fig 1: Example of the interface used by reviewers to review and amend themes



[1] 'Human in the loop' is when human review and decision-making is incorporated into an AI process. This ensures that the outputs are regularly checked, and aims to improve the quality of the process.

All of our code is open source

We have two open source repositories:

- ThemeFinder is a python package containing the AI capabilities (theme generation and mapping). We separated these into their own package to make it easier to test, develop and share with others.
- Consult is our application, containing a UI for both evaluation and for the final dashboard. We plan to continue developing and improving this into a self-serve application.

Research questions

This evaluation was focussed on answering three key questions:

1. **How similarly do Consult and reviewers map themes to responses?**
2. **To what extent do any differences between Consult and the reviewers impact the headline findings of the consultation?**
3. **How could the process be improved?**

Our approach combined quantitative analysis of the Consult and reviewer theme mappings, with user research covering both stages of the Consult process.

Theme mapping quantitative evaluation

Methodology

Once the assigned themes had been reviewed by our subject experts, we had a dataset that included (for each response and question number) the themes that were assigned by Consult, and the themes assigned by our reviewers.

Table 1: Illustrative data snippet (theme names replaced letters, for simplicity)

Respondent	Question #	Consult themes	Reviewer themes
1	1	A, B	A, B, C
2	1	B, D	B, D
3	1	-	-

Research questions

Our main analysis aimed to answer the following questions:

1. **How aligned were the Consult-mapped themes with the reviewer-mapped themes?**
2. **To what extent do differences between Consult and the reviewer themes impact the implied headline findings?**

As well as the themes in the theme framework, we also had two additional options available:

- Other: The response contained themes not covered by the existing categories.
- No Reason Given: The responses gave no clear reason for the views expressed, or the themes were not clearly articulated in the responses.

These were added to help answer the following research questions:

4. **Was the theme framework missing themes that should have been included?**
5. **Can the model identify new themes during the mapping phase?**

Finally, we used data from Consult to answer the question:

6. **How long did it take reviewers to review the Consult themes?**

Consult / reviewer alignment

To assess the alignment between Consult and the reviewer-identified themes, we calculated an *F1 score* – a common measure of a model's performance. Broadly, it considers how often the model correctly identifies a theme (a true positive), versus how often it identifies a theme that is not correct (false positive), or misses a theme that should be there (a false negative). We take the reviewer-identified themes to be the 'correct' answer, and compute the F1 score against this standard[2]. We used a multi-label F1 score that can credit partial matches (where one theme is correct, but another is not), and penalises both false positives and false negatives.

We calculate the F1 score at a response level, and then average across all the responses. For example, row 1 below has an F1 score of 0.8 and row 2 has a score of 1 (perfect alignment), so the average score is 0.9. By calculating the score at a response level and averaging them, longer responses with more themes don't disproportionately drive the score. However, for responses with no identified themes (row 3) it is not possible to calculate an F1 score, so these responses are dropped from the analysis.[3]

Table 2: Example F1 score calculation

Respondent	Question #	Consult themes	Reviewer themes	F1
1	1	A, B	A, B, C	0.8
2	1	B, D	B, D	1
3	1	-	-	-
			Average	0.9

For a fuller explanation of the F1 score and how it is calculated, see the Annex.

Theme ranking

Whilst the number of responses mapped to each theme is important, the qualitative interpretation of the consultation is more driven by their relative frequency (e.g. whether a theme is more prevalent than another) than the absolute numbers. We used the respective theme frequency rankings as a proxy for the expected qualitative interpretation implied by Consult's theme mappings and the reviewers' theme mappings. To understand the extent

[2] In reality, there is nothing saying the reviewer label is 'correct'. Theme mapping is often subjective, and on a previous evaluation (unpublished), two reviewers disagreed on the exact mappings 40% of the time.
[3] We could assign these rows a value of 1, given the Consult and Reviewer themes are the same. However, an F1 score does not consider 'true negatives' like these. While there is an argument that Consult 'correctly' didn't assign any themes, we have excluded these rows to avoid biasing our estimates upwards.

to which the changes made by the reviewers affected these rankings we calculated the difference in rank for each theme. We used these differences to calculate both the 'average distance moved' and the 'maximum distance moved'. The lower these distances are, the more confident we can be that the same qualitative insights would be drawn from the consultation as a whole using either set of mappings.

Use of 'Other' and 'No Reason Given' labels

We summarised the number of times reviewers changed the 'Other' or 'No Reason Given' labels, as well as their overall use by Consult and the expert reviewers. This is complemented by feedback from users to better understand why these labels were used.

Time taken

As the theme mapping evaluation was conducted using the Consult app, we were able to store the timestamp when each response was first seen by a reviewer. We used this to calculate the duration between consecutively reviewed responses to get a proxy for the time taken to check the theme mapping provided by Consult.

Findings

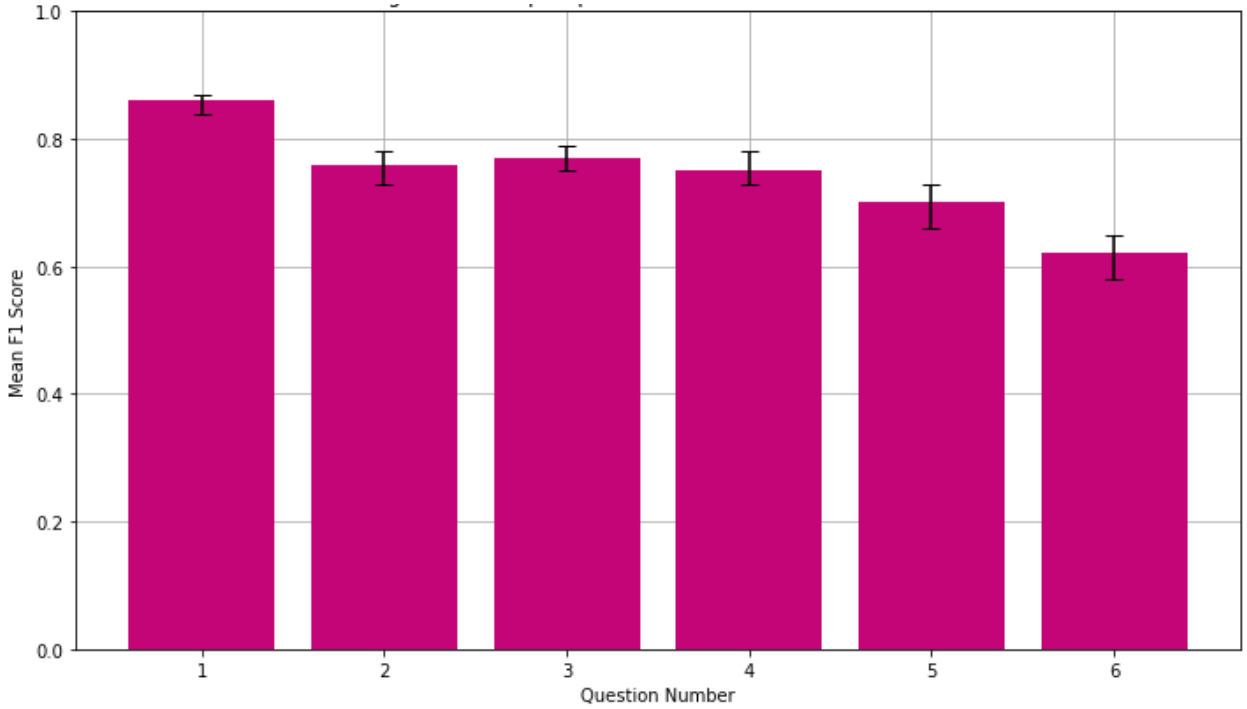
Consult correctly identified all themes for three-fifths of responses and had generally good overall performance, with an average F1 score of 0.76 (out of 1).

For 60% of responses, the themes identified by Consult were unchanged by the reviewers – they were perfectly 'correct'. Calculating performance across all responses (including those with a partial match) the mean multi-label F_1 score was 0.76 out of 1, with a 95% confidence interval of (0.75, 0.77).

There is no set definition of a 'good' F1 score: some sources suggest 0.8-0.9 is good,[4] others over 0.7,[5] and some suggest even lower scores may be considered good in specific contexts.[6] However, 0.76 is close to even the highest definitions for this.

[4] https://encord.com/blog/f1-score-in-machine-learning/
[5] https://spotintelligence.com/2023/05/08/f1-score/#What_is_a_good_F1_score
[6] https://julius.ai/glossary/f1-score

Fig 2: Average F1 scores per question with 95% confidence intervals



This is particularly promising given that there is no objective truth for what a correct theme is – this was based on the assessment of a single reviewer. The potential for subjectivity and human error both make it harder for Consult to be accurate (and mean that even reviewers would not achieve a perfect score). In a previous evaluation of Consult where we had two reviewers, the equivalent F1 score between the reviewers was 0.81 (95% CI: 0.78, 0.83) – only slightly higher than the performance of Consult. Similarly, the reviewers only mapped the exact same themes as each other 62% of the time (roughly the same as Consult's performance on this evaluation).

Differences between Consult and the expert reviewers had minimal effects on the overall theme rankings.

From a policy perspective, identifying the top themes in a consultation is usually more important than knowing the exact number of times the theme appeared. Therefore we are more concerned if the differences in theme mapping lead to the top theme shifting to third place, than if the most common theme is consistent but counted 220 times rather than 300 times.

Comparing the theme rankings resulting from Consult's mapping and those resulting from the reviewers' mapping, we find some differences but they are generally small. For example, looking at Question 2 (chosen as an illustration because it has the fewest themes), no theme moves more than 1 place in the ranking. The Top 3 themes remain unchanged, and only one theme changes in the Top 5.

Table 3: Counts and ranking of each theme, Consult vs. Reviewers (Question 2)

Rank	1	2	3	4	5	6	7	8	9	10
Consult	A (318)	B (277)	C (109)	E (68)	D (61)	F (53)	G (41)	I (16)	H (11)	J (6)
Reviewers	B (329)	A (250)	C (142)	E (74)	F (71)	D (62)	G (57)	I (21)	H (12)	J (10)

While Consult performs particularly well for Question 2, across questions the average distance moved for each theme is generally less than one place. Any movement is primarily driven by a single outlier: each question has, at most, one theme that moves more than two places.

Table 4: Movement of theme rankings, by question

Question number	Number of themes	Average distance moved for each theme	Max distance moved for a theme	Number of themes moving more than 2 places
1	16	0.88	5	1
2	10	0.4	1	0
3	16	1.25	5	1
4	15	0.8	2	0
5	18	0.83	4	1
6	12	1.17	4	1

Reviewers were more likely to add themes than remove them, with changes concentrated on a few key themes.

Inspecting the specific changes made by the expert reviewer, themes were added 1,671 times, but only removed 763 times. Looking specifically at Question 2, removals were heavily concentrated on theme A, which accounted for more than half of all removals. Additions were less heavily concentrated, but theme B was strongly over-represented and accounted for nearly a third of additions. Other questions showed similar patterns.

The fact that differences were driven by a small number of themes may explain the outlier themes that moved substantially in the overall ranking. It also suggests that certain themes are more 'problematic' in terms of how Consult and reviewers interpret them. Future work may enable us to detect and correct these problem themes earlier in the process.

Table 5: Number of times themes were added or removed (Question 2)

	A	B	C	D	E	F	G	H	I	J	**Total**
Added	16	70	35	14	17	29	21	1	8	5	**216**
Removed	84	18	2	13	11	11	5	0	3	1	**148**

For 14% of responses reviewers added an 'Other' theme, suggesting our original theme list was not exhaustive and Consult struggled to make adjustments.

Both Consult and the reviewers were able to label responses as 'No Reason Given' or add 'Other' to indicate a relevant response where the reason was not captured in the current theme list. However, there were substantial differences in how these were used. Both Consult and the reviewers used the 'No Reason Given' label, but this was used 1.5 times more by Consult. In contrast, Consult very rarely assigned the 'Other' label (47 occurrences) while this was used nearly 17 times more by reviewers (nearly 800 occurrences). In one-third (32%) of cases where 'Other' was added, 'No Reason Given' was being removed.

Table 6: Number of times 'No Reason Given' and 'Other' were used, by reviewer

	'No Reason Given'	'Other'
Consult	1086	47
Reviewer	723	793

Reviewers said they selected the 'Other' theme for one of three reasons:

1. They had identified a new theme they thought was important and had not been captured.
2. They needed a way to categorise responses with a negative sentiment.
3. Human error, i.e. creating an 'Other' theme because they had forgotten or missed a pre-existing theme.

While reasons (2) and (3) do not suggest an error in how Consult coded the response, reason (1) suggests that the original theme list was not exhaustive and – crucially – that Consult was not effective at spotting this and suggesting another theme was needed. Participants later reflected that the majority of 'Other' themes had been captured by Consult in the original theme long list (see User research findings). Preserving more of this long list of themes could mitigate this issue.

Reviewing themes generally took little time, but there was some variation to account for response complexity.

The median time taken to review a response was 23 seconds, with 77% of responses reviewed in less than a minute. The time taken to review responses was related to the length of the response and the number of changes reviewers made to the theme mapping. This suggests that while the review process was quick, reviewers did take the time needed to assess longer and more complex answers.

Fig 3: Histogram of time taken to review responses (below 90th percentile)

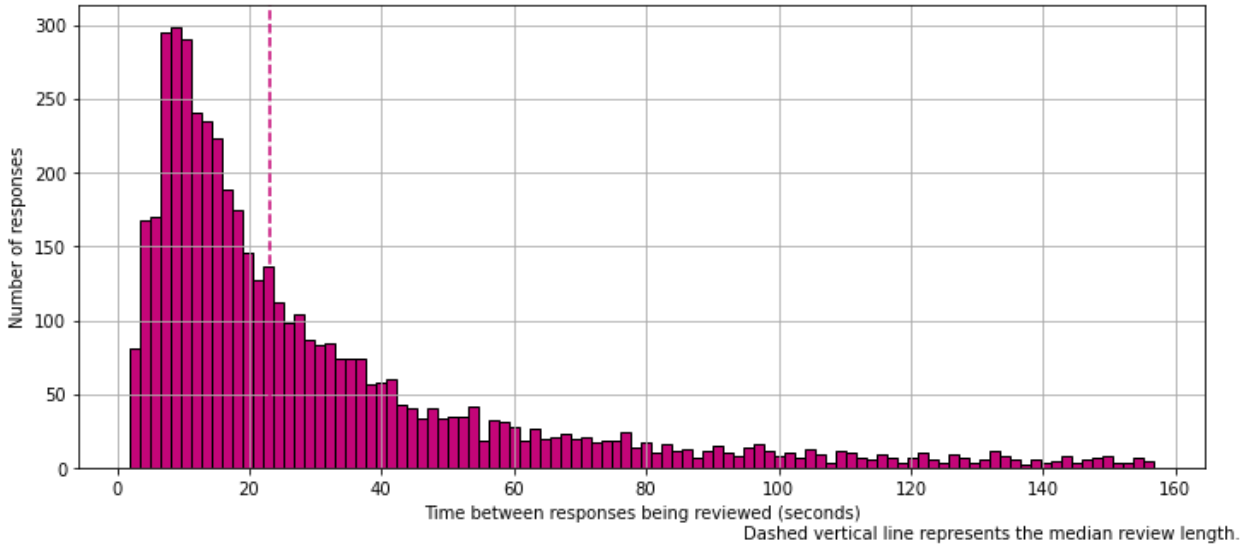


Table 7: Length of response and average edits, by reviewer time

Time	0-5	5-10	10-20	20-30	30-40	40-60	60+
Avg length (words)	11.9	19.6	34.6	47.7	67.8	89.8	138.4
Avg changes	0.07	0.26	0.49	0.81	0.95	0.95	1.10

User research

Methodology

We conducted user research to better understand the experience of the reviewers throughout the process. The majority of this was focussed on two crucial parts of the process:

1. **Theme sign-off.** The process by which expert reviewers can adjust the themes that come out of the theme generation process before they are used in theme mapping.
2. **Theme evaluation.** The process by which the reviewers validate or correct the mappings made by Consult.

A mixed-methods approach was used to understand the behaviour and experiences of the policy staff working on the consultation. This included:

- Observations of the theme sign-off process (n=4)
- Surveys to understand perceptions of the themes and the AI mapping (n=4)
- Interviews and usability sessions to understand experiences (n=5)

Theme sign-off

To evaluate Consult's theme sign-off step, three policy professionals (out of the four involved in this step) gave several rounds of feedback and took part in structured user research sessions.

The first step was to run Consult's theme generation process on all responses to all questions. This generated:

1. A proposed list of themes for each question
2. A longer list of themes that existed before the final condensation step
3. Some example responses for each theme

For each question, the review process proceeded as follows:

1. The assigned policy professional read the short and long list of Consult-generated themes and the example responses.

2. They gave feedback as part of group discussions. For example, they could suggest merging themes, splitting out themes, adding new themes, or editing theme names and descriptions.
3. An AI engineer used this feedback to update the themes, using Consult to ensure the neutral tone of the themes was preserved and to find new example responses.
4. This process was repeated several times until the themes were signed off by all parties.

Findings

The reviewers were impressed with the themes generated.

Overall the reviewers were pleased with the themes generated. They were pleasantly surprised by the first version of themes. Although it took them a lot of effort to go through the theme sign-off process, they were happy with the final set of themes.

> *"[First version of the themes] Pleasantly surprised, useful starting point" (CUR029_1)*

> *"[Final set of themes] Hard to get there but happy with the end result" (CUR029_2)*

The reviewers believed the process reduced bias, however they still valued an opportunity to influence the themes.

The reviewers appreciated the fact that the themes generated by Consult are solely grounded in the responses to the consultation, rather than any pre-conceived expectations of what themes were likely to appear (or that reviewers wanted to see). Compared to the reviewers, Consult may therefore reduce bias in the generation of themes.

> *"Takes away the bias and makes it more consistent...not projecting your own preconceived ideas and when lots of humans are also doing the same, it becomes less and less consistent" (CUR029_4)*

Ultimately, however, the reviewers know what decisions need to be made using the consultation response analysis, and shaping the themes to make sure they get that information is valuable. For example, reviewers split out a theme that captured 'all protected characteristics' into separate themes for each characteristic.

The reviewers wanted more information to assess the validity of the themes.

The reviewers said they needed access to a larger number of consultation responses for each theme, in order to effectively assess whether the theme was sufficient. They initially saw 3 representative responses per theme but this was later changed to 10 based on their feedback. In some cases, participants also needed an explanation of why Consult had mapped certain responses to a theme.

We should consider how to show reviewers more example responses, and experiment with ways to feed their policy questions into the process, for example by starting from the longer list of themes.

The theme sign-off process was cognitively demanding and time intensive.

We found that the process was resource intensive and would not be sustainable for future consultations. Reviewers reported that:

- They had insufficient time with i.AI's engineers but that the session they did have was particularly useful.
- They were unprepared for the amount of time the process would take.
- The process did not get easier as the task progressed and each iteration was as difficult as the last.
- The process was manual and there were a lot of documents to keep track of (each change created a new version of documents, with examples and a long list separately).
- Issues were exacerbated by a 'fear' that because the themes could not be edited after signoff, they had to be perfect.

We should consider different approaches to the sign-off process, including running the mapping on a sample and asking them to review these to allow for more iteration. We should also explore ways to make the information more accessible on a single document.

It is worth noting that without the Consult, creating the theme framework is also a cognitively demanding and time intensive process. This is why the framework is often only created on a sample of the data.

After completing the evaluation of the mapping phase, the reviewers thought they should have trusted the original AI themes more and changed them less.

In the quantitative analysis, we observed that reviewers often added the 'Other' label to responses (see here), indicating that there was a reason given that was not captured in the current themes. However, when these responses were revisited by reviewers, they often felt that the response had been captured by the long list of themes and felt they could have been more confident in the original list of themes.

> *"Looking back at the initial theme long list, having more confidence in the initial AI proposed themes...may have made the theme checking more logical" (CUR029_3)*

Overall, they recommended that other departments using Consult should have more trust in the themes Consult generates and avoid too many changes to themes, as they felt like they had spent most of their time tweaking around the edges. Some of that was driven by uncertainty of what kinds of changes are important to make, which is something we could provide more clarity on.

By reducing the time needed for analysis, Consult allowed them to focus more on the 'so what'.

It was clear that the use of Consult, both in the generation of the themes and the mapping stage, sped up the process of analysing responses and allowed them to focus more on interpretation. This finding is in line with the quantitative analysis (see here), showing how quickly they were able to review Consult's mappings.

> *"Use of AI incredibly useful...sped up this phase, got to the outcome sooner so we can get to the analysis and draw out what's needed next" (CUR029_3)*

> *"You've done grand...easy to do a sense check...saved us a heck of a lot of time" (CUR029_4)*

The reviewers would prefer a process where they didn't check 100% of responses, but raised concerns about trusting the outputs with no human checks.

Participants reflected that they still think some form of human in the loop is needed to fully trust the outputs, but they also need to save time and would prefer not to need to review all responses.

"As a user of this tool it would be great not to have to do 100% checking...can we use the AI to get the mapping better next time round" (CUR029_1)

We should explore approaches to iteratively improving the themes during the Consult process to improve the mapping performance. For example, by improving the theme description and labels to remove ambiguity.

Areas for improvement

We identified the following areas for improvement:

- **Experiment with new ways of presenting the data needed for sign-off**, including:
  - Starting with a longer list of themes and providing tools to make it easier to whittle them down.
  - Ensuring all the relevant information is in one place and not spread across several documents.
- **Experiment with running the mapping as part of the theme sign-off process**. This will not only give the reviewers a better understanding of how the themes will be mapped, but allows them to spot any themes that are being consistently mapped incorrectly.
- **Better communicate the overall process** so that partners can better plan their resource allocations and participants will have a greater contextual understanding of individual stages and what changes are important.

Conclusion

Consult is an effective tool for analysing consultations:

- It identifies good themes in the data.
- Variation between Consult and reviewers was similar to the expected level of person-to-person difference.
- Differences between Consult and reviewer-identified themes had negligible impact on the overall theme rankings, implying similar qualitative interpretations.
- Consult speeds up analysis and allows the reviewers to focus more on the 'so what'.

However, there are areas for improvement:

- There were discrepancies between Consult and the reviewer-mapped themes. While this is partly due to the subjective nature of the task, it can impact reviewers' trust in the process. The ability to review and alter theme mappings through the Consult app is important for maintaining that trust.
- It is hard to benchmark Consult's performance, given theme mapping is inherently subjective. This would be easier if responses were double reviewed, which we will test in future evaluations.
- Consult rarely flags new themes (using the 'Other' theme) during the mapping stage. It is therefore important that the correct themes end up in the framework.
- The current theme sign-off process is cumbersome and would not scale well.

Focuses for future evaluations:

In our future rounds of evaluation, we will prioritise the following:

- ➔ Refining the theme sign-off process, focusing on usability and ensuring the correct themes end up in the framework. This could include mapping a small sample as part of the sign-off process.
- ➔ Testing whether amending specific themes can reduce the discrepancy between Consult and reviewers.
- ➔ Using two reviewers per response to benchmark Consult against person-to-person alignment.

Annex

This annex provides a more detailed explanation of the multi-label F_1 score that we used to assess the performance of Consult relative to the reviewers.

To clarify the definition, we can first define some common terms using the classic 2×2 confusion matrix setup.

- Note that we're only using the simpler 2×2 case to make our initial definitions clearer. In the actual analysis we also extended the metric to the multi-label case.

		Predicted category		
		Positive	Negative	
Actual Category	Positive	True Positive (TP)	False Negative (FN)	Σ=Actual Positives (AP)
	Negative	False Positive (FP)	True Negative (TN)	Σ=Actual Negatives (AN)
		Σ=Predicted Positives (PP)	Σ=Predicted Negatives (PN)	

Building on the 2×2 matrix above, we can define eight ratios that help us characterise the performance of our classifiers (i.e. the output from the reviewers and Consult):

True Positive Rate $TPR = \frac{TP}{AP}$	False Negative Rate $FNR = \frac{FN}{AP}$
• The proportion of actual positives that are correctly predicted to be positive. • $TPR = 1 - FNR$ • A.k.a. Recall, Sensitivity	• The proportion of actual positives that are incorrectly predicted to be negative. • $FNR = 1 - TPR$ • A.k.a. Type II error
True Negative Rate $TNR = \frac{TN}{AN}$	False Positive Rate $FPR = \frac{FP}{AN}$
• The proportion of actual negatives that are correctly predicted to be negative. • $TNR = 1 - FPR$ • A.k.a. Specificity	• The proportion of actual negatives that are incorrectly predicted to be positive. • $FPR = 1 - TNR$ • A.k.a. Type I error
Positive Predictive Value $PPV = \frac{TP}{PP}$	False Discovery Rate $FDR = \frac{FP}{PP}$
• The proportion of predicted positives that are actually positive. • $PPV = 1 - FDR$ • A.k.a. Precision	• The proportion of predicted positives that are actually negative. • $FNR = 1 - TPR$
Negative Predictive Value $NPV = \frac{TN}{PN}$	False Omission Rate $FOR = \frac{FN}{PN}$
• The proportion of predicted negatives that are actually negative. • $NPV = 1 - FOR$	• The proportion of predicted negatives that are actually positive. • $FOR = 1 - NPV$

F_1 score

The F_1 score combines two of the eight ratios from the table above, True Positive Rate and Positive Predictive Value, to give an aggregated measure of performance in relation to the 'positive' class. Specifically, F_1 score is the harmonic mean of the True Positive Rate (TPR) and the Positive Predictive Value (PPV):

$$F_1 = \frac{2}{TPR^{-1} + PPV^{-1}}$$

During our evaluation, more than one theme could be mapped to a given response. This meant that we were evaluating a multi-label classification problem and needed to adjust our metric accordingly.

To do this, we converted the theme mappings into sparse matrices where each column represented one of the theme labels that could have been chosen and each row represented a consultation response. We then calculated the F_1 score for each individual response and averaged these scores to get the aggregate values reported. The benefit of this approach is that each response contributes equally to our assessment of performance.

This process is illustrated in the hypothetical below:

- Imagine that there were four possible themes (A, B, C, D), and for a particular response we received the following mappings from our labellers:
    - Reviewer = (A, C)
    - Consult = (A)
- We would convert the data into sparse matrices with a column per possible topic:
    - Reviewer = (1, 0, 1, 0)
    - Consult = (1, 0, 0, 0)
- We would calculate the True Positive Rate and Positive Predictive Value for this response:
    - TPR = 0.5
        - There are two ‘actual positives’ in the reviewer’s answer.
        - Consult has correctly ‘predicted’ one of them.
    - PPV = 1
        - There is one ‘predicted positive’ in Consult’s answer.
        - This predicted positive is ‘correct’ according to the reviewer.
- We would calculate the F_1 score for this response using the formula above:

- $F_1 = \frac{2}{TPR^{-1} + PPV^{-1}} = \frac{2}{\frac{1}{0.5} + \frac{1}{1}} = \frac{2}{3}$

- This process would be repeated for all the responses and the arithmetic mean of the response-level F_1 scores would be calculated.

We calculated this statistic using a function from scikit-learn's metrics module: f1_score().

Confidence intervals

In addition, we generated a confidence interval for the F_1 score using a process called bootstrapping. This involved taking a large number of random samples of responses from the data and calculating the F_1 score for each of these samples. This process gave us an approximate sampling distribution for the metric itself. Confidence intervals can be extracted from this sampling distribution by finding the relevant percentiles. We calculated 95% confidence intervals by finding the 2.5th and 97.5th percentiles. The confidence interval gives us a sense of the uncertainty we have in our metric.