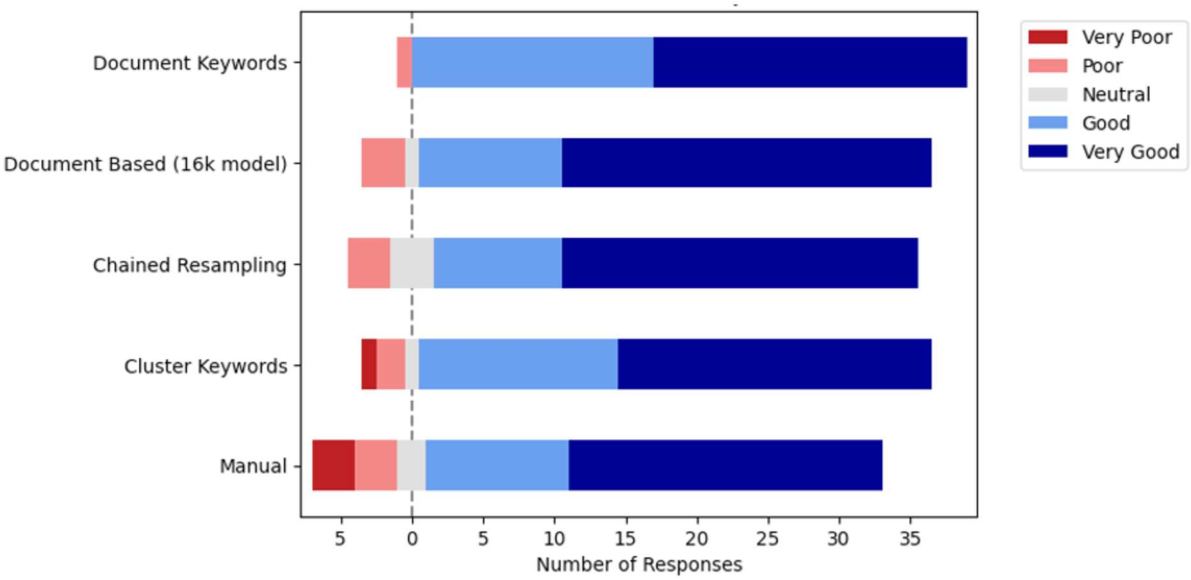
Supplementary Material

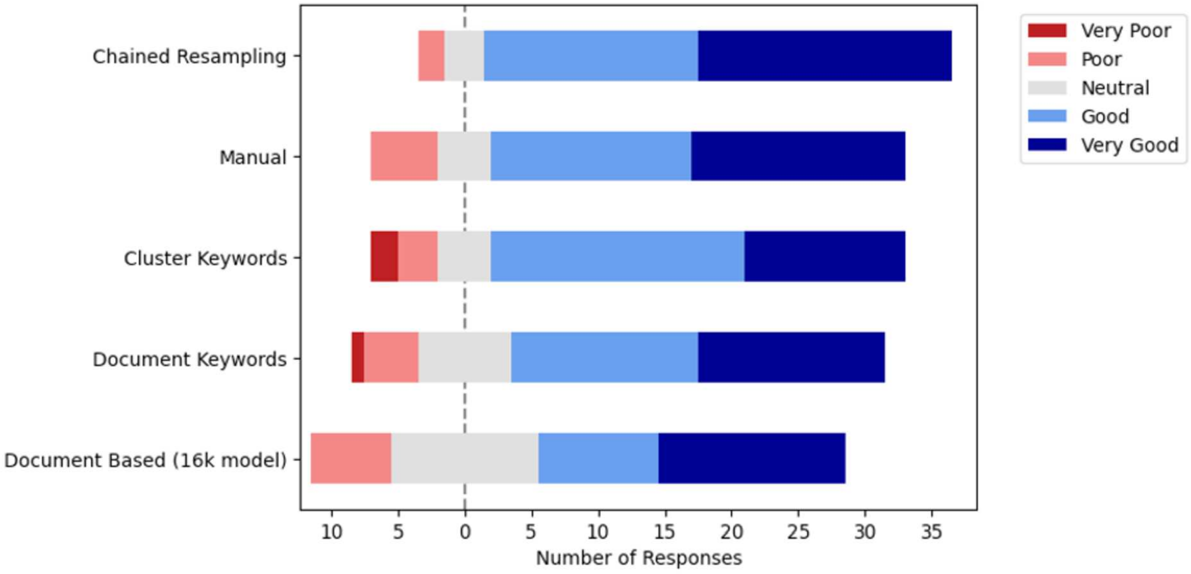
Supplemental Figures 1–13, Supplemental Tables 1–5.

A Supplementary Figures

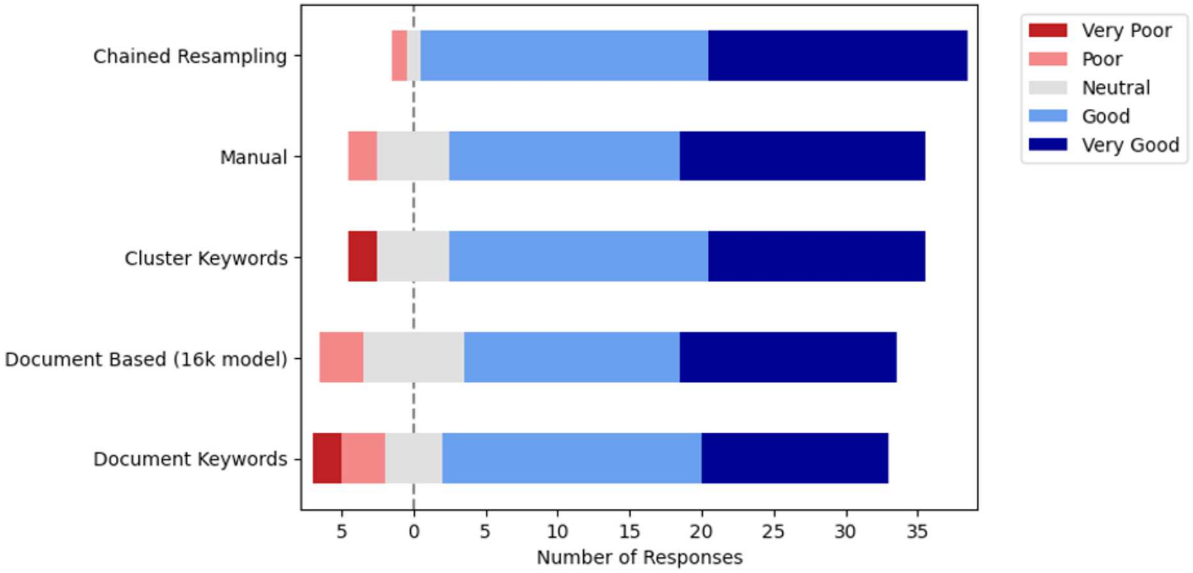
This section contains the Supplementary Figures referenced throughout the main text.



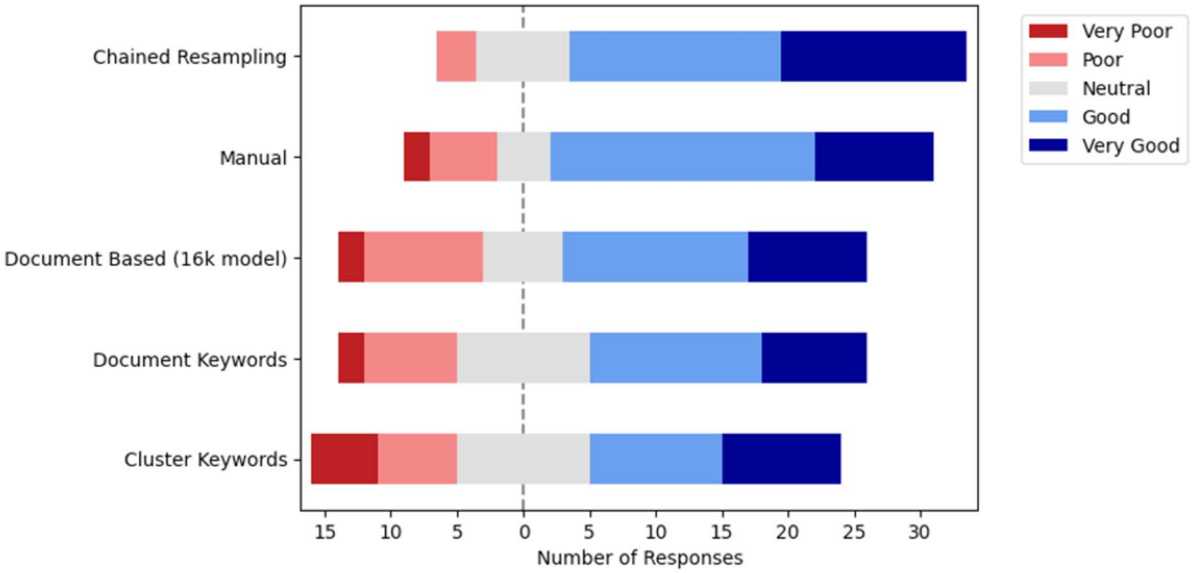
Supplemental Figure 1: NSF name fluency.



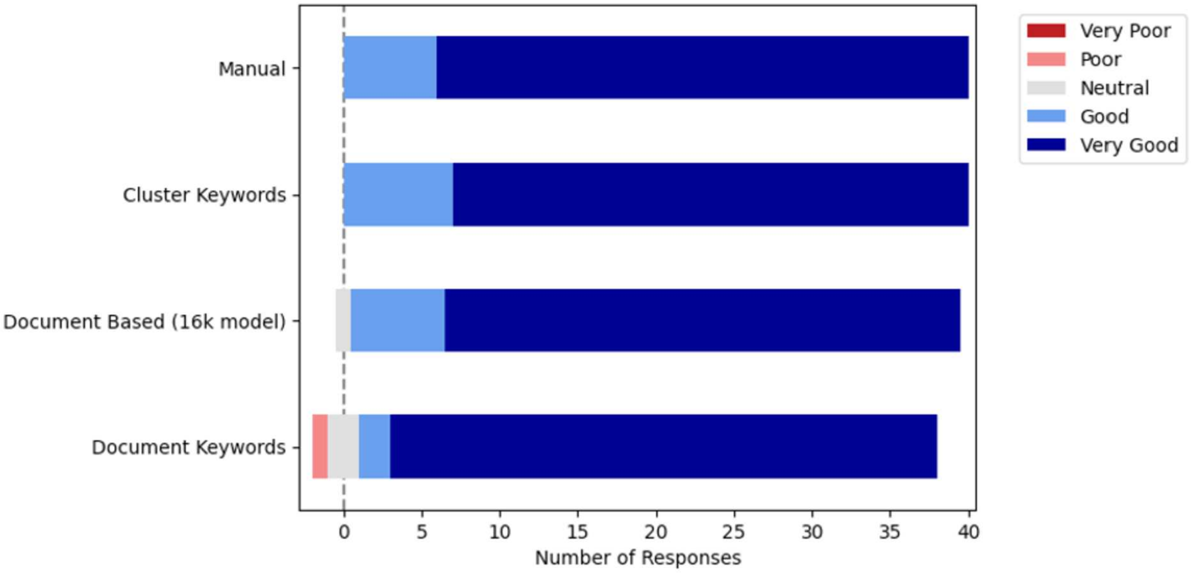
Supplemental Figure 2: NSF name consistency.



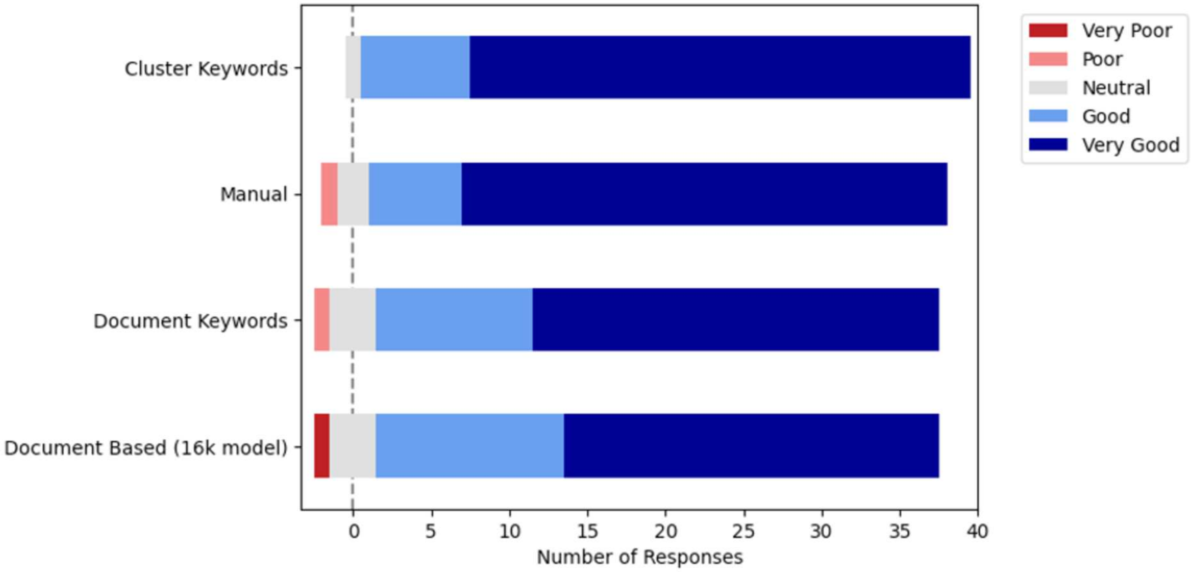
Supplemental Figure 3: NSF name relevance.



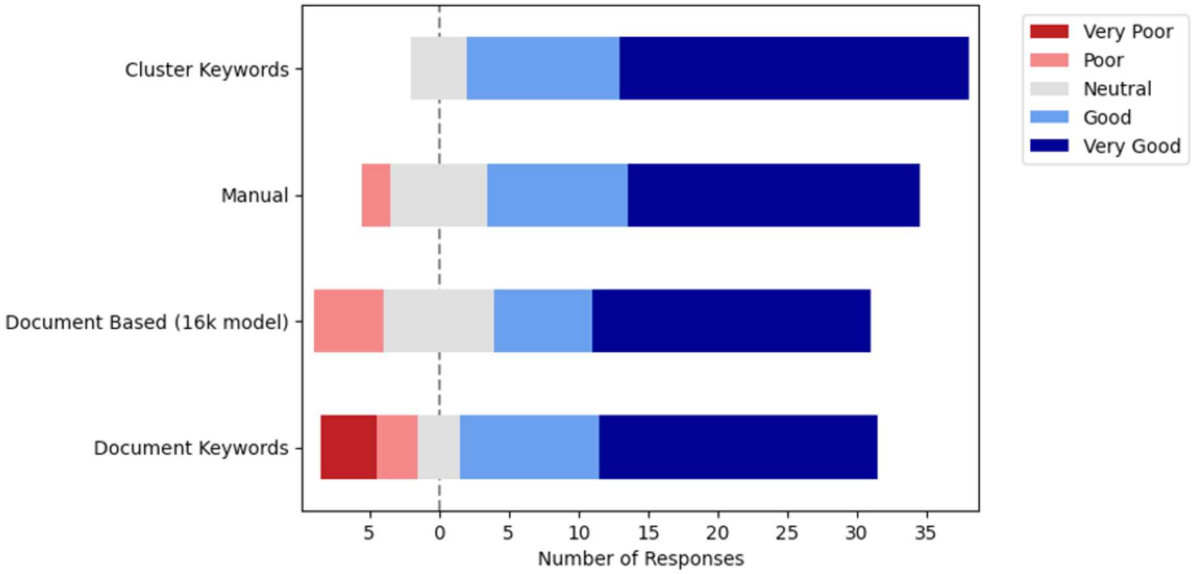
Supplemental Figure 4: NSF name completeness.



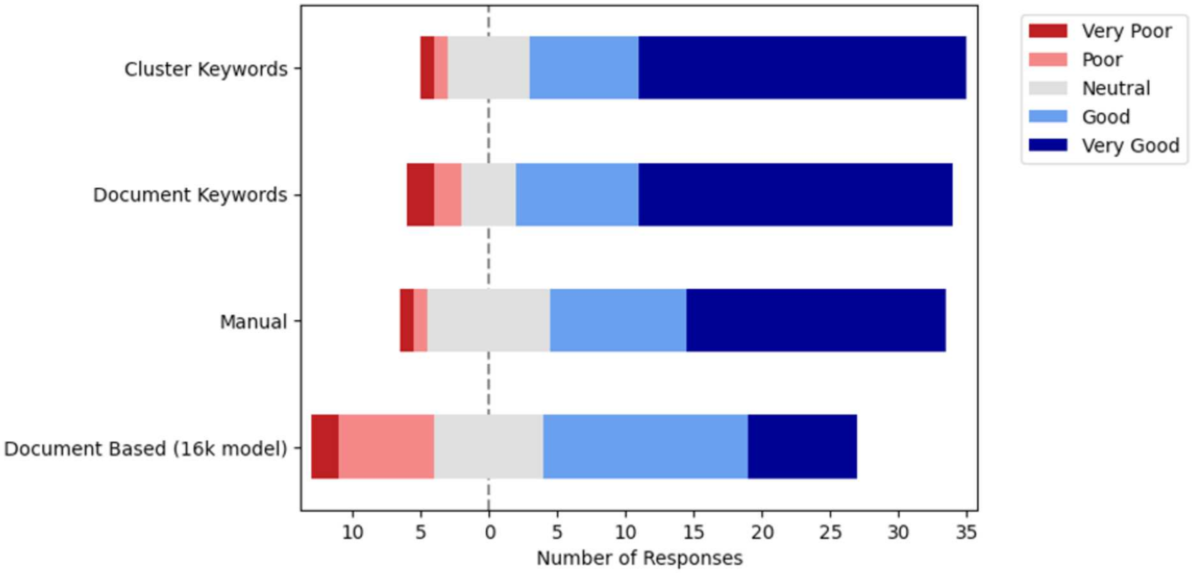
Supplemental Figure 5: PMC name fluency.



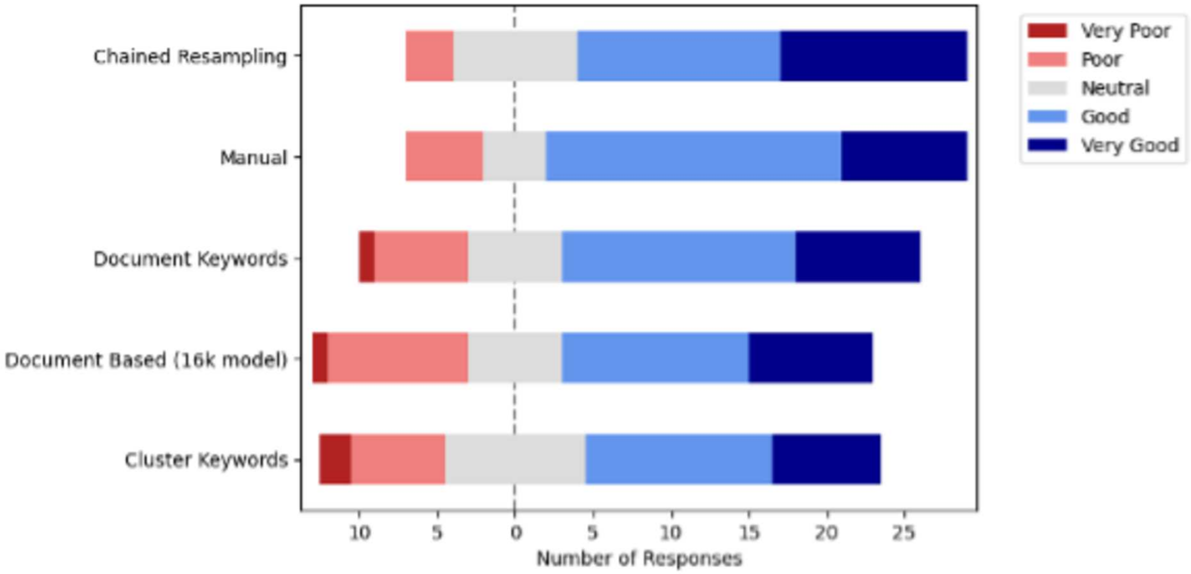
Supplemental Figure 6: PMC name consistency.



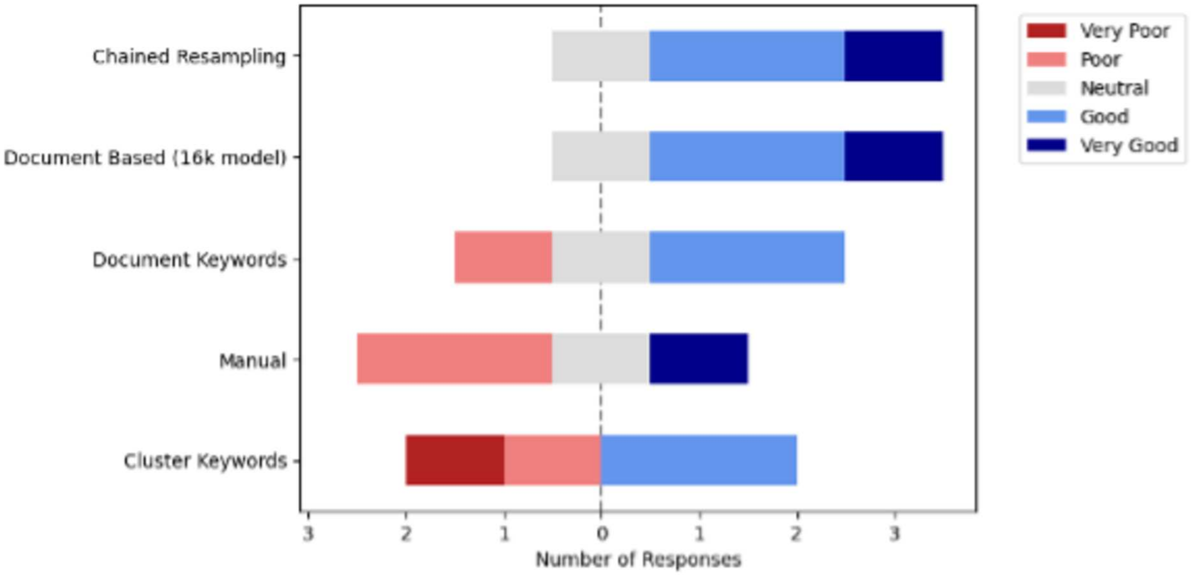
Supplemental Figure 7: PMC name relevance.



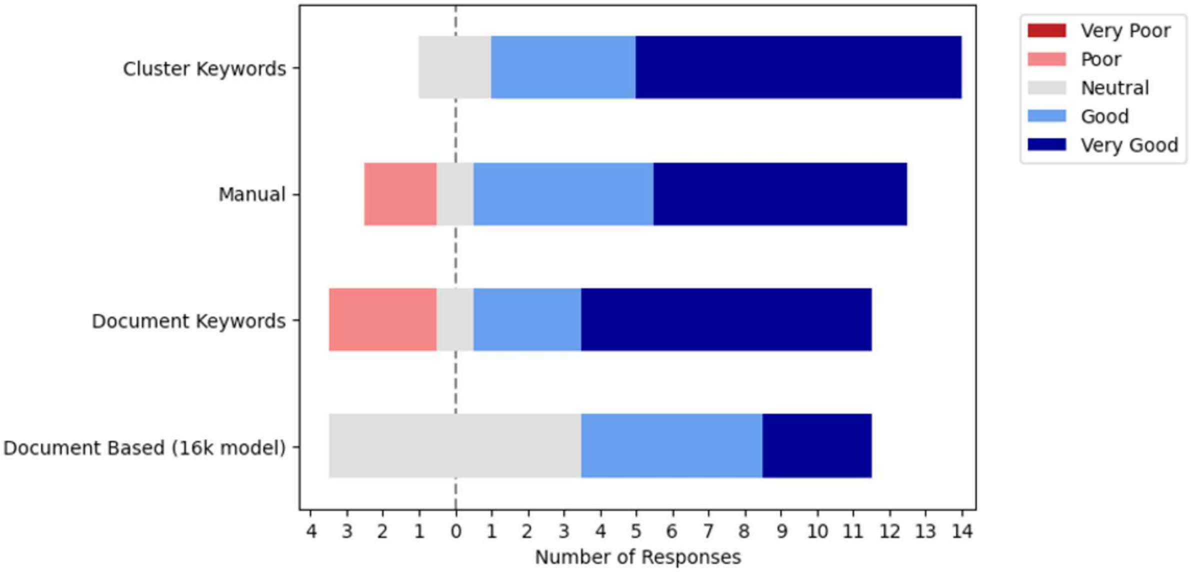
Supplemental Figure 8: PMC name completeness.



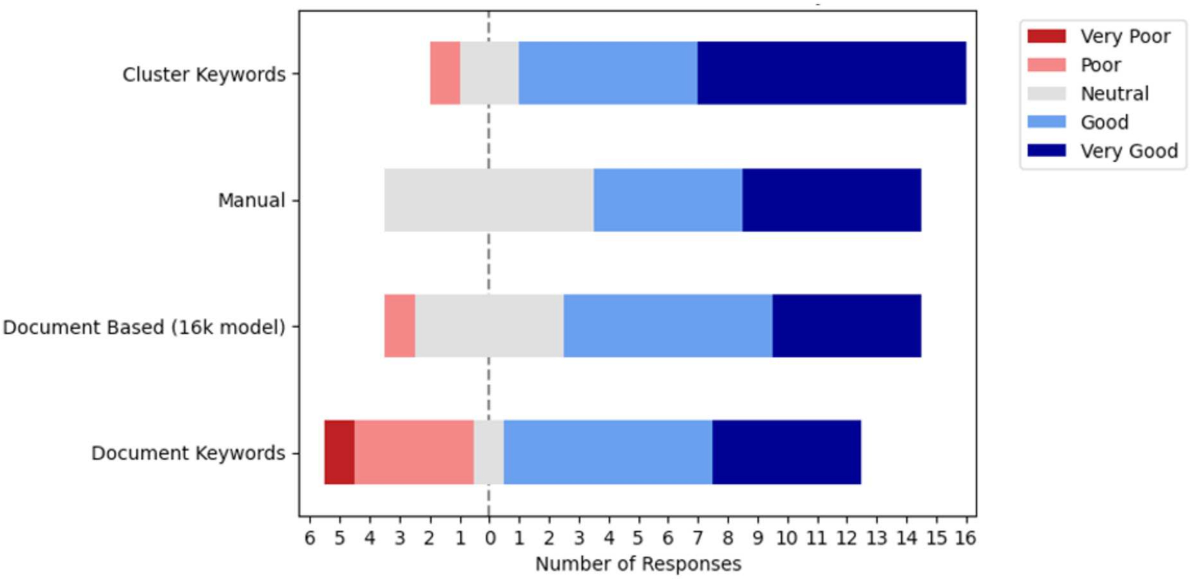
Supplemental Figure 9: NSF overall name quality, cluster size <100.



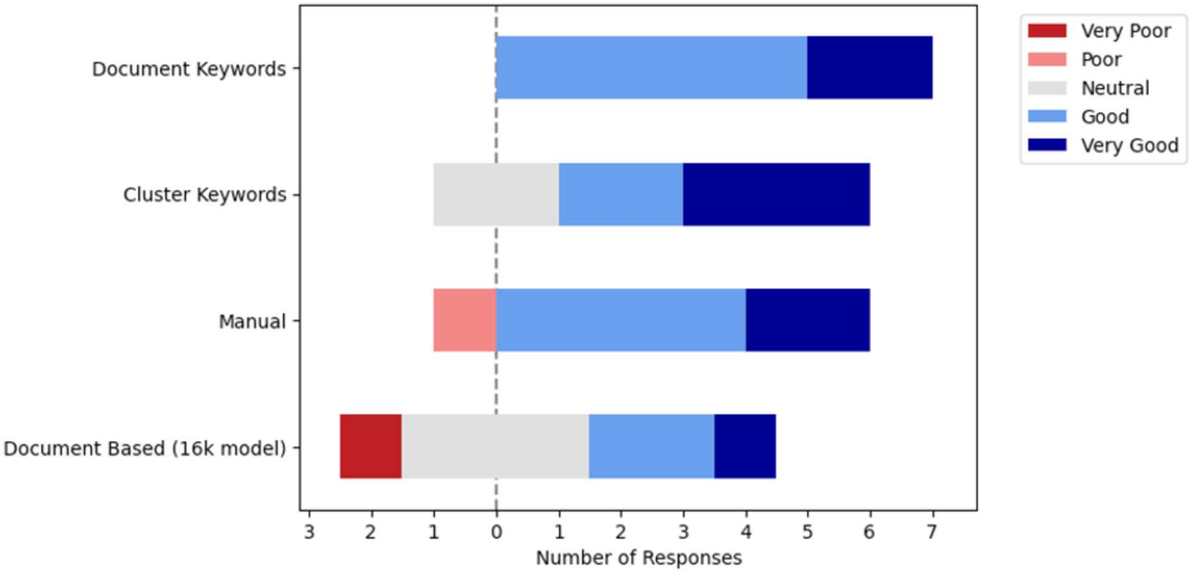
Supplemental Figure 10: NSF overall name quality, cluster size 100–500.



Supplemental Figure 11: PMC overall name quality, cluster size <100.



Supplemental Figure 12: PMC overall name quality, cluster size 100–500.



Supplemental Figure 13: PMC overall name quality, cluster size >500.

B Supplementary Tables

Supplemental Table 1: Prompts.

Naming Method	Prompt Text
Document-based	“Please identify the common topic/theme among the following texts, being as precise as possible. Disregard any outliers. Provide only the topic name.”
Document Keywords and Cluster Keywords	“Please identify the common topic/theme among the following set of keyphrases, being as precise as possible. Disregard any outliers. Provide only the topic name.”
Chained Resampling	“The following label candidates have been generated for a cluster of documents. Each label candidate will be enclosed in brackets. Evaluate the candidates and generate a consensus label that best represents them. Respond with only the text of the consensus label.”

Supplemental Table 2: Cluster quality by cluster size, NSF abstracts dataset. Chi-squared test statistic $= 6.48$, $p = 0.166$, Cramer’s $V = 0.36$.

Cluster Quality	Very Poor	Poor	Neutral	Good	Very Good
Cluster Size Bin					
<100	3 (6.98%)	2 (4.65%)	2 (4.65%)	22 (51.16%)	14 (32.56%)
100–500	0 (0.00%)	1 (14.29%)	2 (28.57%)	3 (42.86%)	1 (14.29%)

Supplemental Table 3: Cluster quality by cluster size, PMC patients dataset. Chi-squared test statistic = 10.9, $p = 0.0922$, Cramer's $V = 0.333$.

Cluster Quality	Very Poor	Poor	Neutral	Good	Very Good
Cluster Size Bin					
<100	0 (0.00%)	0 (0.00%)	4 (21.05%)	4 (21.05%)	11 (57.89%)
100–500	0 (0.00%)	0 (0.00%)	2 (10.00%)	7 (35.00%)	11 (55.00%)
>500	0 (0.00%)	2 (20.00%)	1 (10.00%)	4 (40.00%)	3 (30.00%)

Supplemental Table 4: Raw coded cluster names. This table is too large to appear here. See supplementary file, "supplementary Table 4.csv" or the Supplementary Table 4 sheet of supplementary file "all tables and supp tables.xlsx".

Supplemental Table 5: Cost comparisons for model-generated and human-generated names.

Prompting Strategy	NSF Abstracts		PMC Patients	
	Per-Cluster Cost	Total Cost (123 clusters)	Per-Cluster Cost	Total Cost (274 clusters)
Document Keywords*	$0.00	$0.18	$0.00	$0.55
Cluster Keywords*	$0.00	$0.08	$0.00	$0.21
Document-Based*	$0.13	$15.88	$0.13	$36.21
Chained Resampling*^	$0.05	$6.14	–	–
Manual Naming – Lower Bound ($4/hour at 10 minutes per cluster)	$0.67	$82.41	$0.67	$183.58
Manual Naming – Upper Bound ($50/hour at 25 minutes per cluster)	$20.83	$2,562.50	$20.83	$5,708.33

C Guide to Cluster Naming and Cluster Name Evaluation

See supplementary file "Supplemental Section 1 – Annotation Guide.pdf"

D Blueprint for Using LLM Cluster Naming on Your Project

See supplementary file "Supplemental Section 2 – Blueprint.pdf"