Review by Chat GPT 2026 09 05

Please examine this one page web application, which i developed with the extensive assistance of a LLM AI:
https://mande.co.uk/wp-content/uploads/2026/05/20260903_TextClusterAnalysisLab_Vs136.html

I would like your independent opinion on its strengths and weaknesses. An introduction has been given the app elsewhere which might also provide more context, in addition to the Introduction tab on the app itself:

https://mandenews.blogspot.com/2026/05/cluster-analysis-for-evaluation.html

and

https://mande.co.uk/special-issues/the-text-cluster-analysis-lab/

Yes. I have examined the three materials you supplied, including the current **Vs136 (3 September 2026)** description and the methodological discussion around it.

One limitation is worth making explicit: the standalone HTML application itself is not exposing its interface/code to my web reader, so I cannot yet conduct a true *interface-level/code-level audit* of Vs136. I can, however, give you an independent assessment of the **method, workflow, conceptual design, documentation, and likely strengths/weaknesses of the application as described and demonstrated through the surrounding material**.

My overall judgement is quite positive:

> **I think this is a genuinely interesting methodological contribution, rather than merely an LLM-assisted convenience tool. Its strongest feature is that it makes the *basis of the clustering inspectable*. Its greatest vulnerability is that the apparently objective binary criteria are themselves generated and applied by an LLM, so the method's reproducibility and validity ultimately depend on how well that layer is controlled.**

1. The central idea is strong

The most important design decision is this:

LLM → human-readable criteria → binary matrix → conventional clustering

rather than:

LLM/embedding model → opaque numerical representation → clustering → LLM explains result

That distinction is important.

The Lab makes the *features that determine the clustering* visible before the clustering takes place. Your description correctly emphasises that the dendrogram's explanations are therefore not simply post-hoc interpretations.

That gives the method an unusually good **chain of interpretability**:

> corpus → analytical frame → criteria → criterion scoring → binary matrix → similarity → dendrogram → interpretation

This is, in my view, the application's strongest intellectual contribution.

It also makes the method particularly appropriate for evaluation, where the evaluator generally needs to be able to answer not just **"What groups did you find?"**, but **"Why are these cases grouped together, and could somebody else inspect the reasoning?"**

---

2. The application is unusually good at exposing analytical choices

This is another major strength.

The Lab doesn't pretend that clustering is an automatic discovery process. It exposes choices such as:

- what the texts are;
- what kind of difference is being sought;
- which criteria are retained;
- how reliability is assessed;
- which clustering measure is used;
- how clusters are linked;
- where the dendrogram is cut;
- what questions are subsequently asked of the groups.

Your accompanying discussion makes this point particularly well: choices can be **visible, checkable, self-signalling and documented**, rather than hidden inside an algorithm.

I think that is more important than it may initially appear.

Many quantitative clustering procedures give users a large number of technically sophisticated options while giving relatively little assistance in understanding **which decisions are substantively consequential**.

Your Lab reverses that emphasis.

It is saying, in effect:

> *The analyst should be able to see the conceptual machinery that produced the classification.*

That is very compatible with evaluation methodology.

---

3. The binary matrix is a particularly good design choice

This is probably the most underrated part of the application.

The dendrogram is visually attractive and useful, but the **text × criterion matrix** is what gives the method analytical substance.

Once you have:

Text	Criterion A	Criterion B	Criterion C	Criterion D
T1	1	0	1	1
T2	1	1	0	1
T3	0	0	1	0

you have moved from an LLM conversation into a structure that can be inspected, exported, analysed and independently recomputed.

That is a substantial methodological advantage.

It also means the Lab is not really "AI clustering" in the usual sense. The LLM is being used primarily as a **measurement/coding instrument**, while the actual clustering is performed using established mathematical procedures.

That distinction should, in my opinion, be made even more prominent in the application's methodological description.

---

4. The reliability idea is clever — but this is also where I see a significant weakness

Your "back-translation" reliability test is an ingenious response to an obvious problem:

> How do we know that Claude would apply its own criterion consistently?

Having another pass effectively reconstruct whether the criterion can reproduce the classifications is a sensible way of probing **coding reliability**.

However, I would be careful about the word **reliability**.

The test establishes something like:

> *How consistently can the criterion be operationalised by the model?*

It does **not** establish:

> *Whether the criterion is substantively valid.*

You actually acknowledge this in the accompanying article, and I think that qualification is extremely important.

There is a deeper issue, though.

If the **same LLM** generates a criterion and then performs the reliability test, the test is not fully independent.

It is rather like asking a human researcher:

> "You created this coding rule. Now apply it again. Did you apply it consistently?"

That tells us something, but less than an independent coder would.

I would therefore consider adding a stronger form of reliability test

For example:

Criterion generated by Model A → scored by Model B

or:

Criterion generated by Claude → scored independently by Gemini

or, ideally:

LLM scoring → human sample audit

That would turn your current reliability mechanism into something much more defensible methodologically.

You could perhaps distinguish:

- **internal LLM consistency**
- **cross-model agreement**
- **human–LLM agreement**

That would be a very interesting methodological experiment in its own right.

---

5. The biggest methodological weakness: the criteria are doing almost everything

This is the point I would emphasise most strongly if you were preparing this for methodological scrutiny.

You correctly say that the criteria determine the matrix and that everything downstream follows from that matrix.

That means:

> **criterion generation is effectively the measurement model.**

The clustering algorithm is comparatively uncontroversial.

If I give two analysts exactly the same binary matrix, the differences between their dendrograms are largely a matter of technical choices.

But if I give the LLM two different prompts:

> "Identify differences in outcomes"

versus

> "Identify differences in the participation of marginalised groups"

I can obtain dramatically different matrices — and therefore dramatically different dendrograms.

This is not necessarily a flaw. In fact, I agree with your argument that different framings producing different trees can be informative.

But it means that the Lab should perhaps be understood less as:

"a method for discovering the structure of a corpus"

and more as:

"a method for exploring the structure of a corpus under explicitly specified analytical lenses."

That is a subtle but important distinction.

---

6. I think your strongest methodological claim needs slightly more restraint

I would be cautious about claiming that the Lab has solved the "subjective naming" problem.

It has certainly **reduced** the problem.

The dendrogram can report:

> Cluster A differs from Cluster B because A contains criteria X and Y while B contains criteria Z.

That is excellent.

But somebody still has to decide that:

> X is an important distinction.

And somebody has to decide that the criterion itself is a meaningful representation of the underlying text.

So I would describe the achievement as:

"making cluster interpretation inspectable"

rather than:

"removing subjectivity from cluster interpretation."

Your own discussion actually reaches this more nuanced position, and I think it is the more defensible one.

---

7. The Selection tab may ultimately be more important than the dendrogram

This is something I found particularly interesting in the current description.

The addition of Selection turns the application from:

> "Here is a clustering generated from these criteria"

into:

> "Here are alternative ways of operationalising the distinctions in these texts; now examine their reliability, discrimination and overlap and construct a defensible

> criterion set."

That is a much more sophisticated analytical proposition.

It also brings the Lab much closer to **feature selection** and **measurement design** than ordinary text clustering.

Your proposed next step — examining the relationship between:

number of criteria ↔ discriminating power

— could be especially valuable.

In fact, I would make that a major future development.

A graph such as:

cumulative discriminating power vs number of criteria

would allow the analyst to see whether the extra criteria are producing substantial additional differentiation or merely marginal gains.

That would give a concrete empirical basis for discussing **parsimony**.

---

8. I would strengthen the comparison with conventional thematic analysis

The distinction you make between thematic coding and your approach is broadly persuasive:

- thematic analysis generally asks **what themes occur within texts?**
- the Lab asks **how do whole texts differ from one another?**

That is an important distinction.

But I would avoid making conventional thematic analysis sound quite as homogeneous as the current description sometimes does.

There are forms of qualitative comparative analysis, framework analysis, typological analysis, qualitative content analysis, and structured case comparison that are already concerned with **differences between cases**.

So your most defensible claim is not:

> "Existing qualitative analysis cannot do this."

It is:

> **"The Lab provides a particularly explicit and computationally reproducible way of constructing a cross-text typology while retaining human-readable criteria."**

That is still a strong claim — and harder for a methodological critic to knock down.

---

9. There is an important issue with the word "cluster"

I would seriously consider this.

The Lab produces a **classification based on selected binary attributes**.

That is certainly clustering in the broad sense.

But the resulting groups are not necessarily "natural clusters" inherent in the corpus.

They are clusters **relative to the selected feature space**.

This matters because a sophisticated user might ask:

> "Are these genuinely occurring groups of texts, or are they artefacts of the criteria supplied to the algorithm?"

The answer is essentially:

both possibilities need to remain open.

The dendrogram is telling you:

> "Given these criteria and this similarity measure, this is the resulting structure."

That is perfectly legitimate.

But I would make that conditional nature more explicit.

---

10. The Query function is powerful — but potentially dangerous

I like this feature very much.

It changes the dendrogram from being merely a visualisation into an **analytical interface**.

The ability to ask:

> "How does this group differ from its sibling?"

is particularly well aligned with the hierarchical structure.

But there is a conceptual boundary that users could easily misunderstand:

The Query function is not constrained by the matrix.

Your documentation does now explicitly warn about this.

That warning is essential.

Otherwise someone could obtain:

1. a statistically/mathematically defined split based on criteria A, B and C;
2. ask Claude why the groups differ;
3. receive an answer mentioning D, E and F;

and assume D, E and F somehow contributed to the clustering.

They did not.

I would therefore make the visual distinction between:

Matrix-derived finding

and

LLM-generated interpretation

extremely strong in the interface.

---

11. Provenance is one of the Lab's outstanding strengths

The exportable provenance information is excellent.

The fact that you retain:

- prompts;
- model;
- temperature;

- criteria;
- reliability results;
- clustering parameters;
- token usage;
- costs;
- query dialogue;
- session state;

means that the application has something many AI-assisted research tools badly lack:

an audit trail.

The JSON session facility is particularly valuable because it potentially allows somebody else to reproduce the *analytical state*, rather than merely receiving a PDF containing the final results.

I would regard this as a major selling point.

---

12. The application's biggest practical weakness may actually be cognitive overload

From the surrounding description, I infer that Vs136 has accumulated quite a lot of functionality:

- setup;
- criteria generation;
- scoring;
- reliability;
- selection;
- clustering;
- queries;
- exports;
- provenance;
- multiple parameters.

That is analytically impressive.

But there is a danger that the application becomes something that **you understand very well because you built it, but a first-time evaluator does not.**

This is probably the biggest UX risk.

The methodological complexity is legitimate.

The interface should therefore distinguish very clearly between:

Essential decisions

1. What are the texts?
2. What kind of difference are you interested in?
3. Which criteria do you accept?
4. Are the criteria reliable?
5. What does the resulting classification show?

and:

Advanced decisions

- distance measure;
- linkage method;
- temperature;
- model selection;
- sample size;
- cut level;
- etc.

I would hide or collapse many of the latter under **Advanced settings**.

---

13. I would add one feature above almost everything else

If I were independently advising you on the next version, I would prioritise:

"Why is this text in this cluster?"

For every text, let the user click it and see something like:

> **Text 17**
>
> Assigned to Cluster 4 because it shares:
>
> - Criterion A
> - Criterion C
> - Criterion F
> - Criterion H
>
> with its nearest neighbours.
>
> It differs from Cluster 5 principally because Cluster 5 has:

- Criterion B
- Criterion D

That would make the entire classification auditable **at the individual-case level**.

You already have much of the underlying information to do this.

---

14. A second feature I would strongly recommend

I would also create a **"sensitivity analysis"** function.

For example:

> Run 1 — criteria set A
> Run 2 — criteria set B
> Run 3 — criteria set C

Then show:

Which relationships between texts remain stable?

This would address one of the most important epistemological questions:

> *Which findings are robust to reasonable alternative operationalisations?*

That is potentially much more informative than presenting a single "best" dendrogram.

It would also fit extremely well with your argument that different framings deliberately produce different structures.

---

15. The MSC example is particularly valuable

The example where the Lab failed to predict the panel's choices is, paradoxically, one of the strongest parts of the whole presentation.

You report that the Lab predicted two different stories from those actually selected by the human panel, and then used the discrepancy to expose two different conceptions of "significance."

That is much more interesting than saying:

> "The AI successfully reproduced the experts."

The disagreement demonstrates that the tool can become an **object of evaluation itself**.

That suggests a potentially important use of the Lab:

> **not to replace human judgement, but to make differences between human and machine judgement analytically visible.**

That is a very strong proposition for evaluation.

---

16. My overall assessment

If I were reviewing this as a methodological prototype, I would score it approximately as follows:

Dimension	My assessment
Originality of concept	**9/10**
Interpretability	**9/10**
Transparency	**9/10**
Evaluation usefulness	**8.5/10**
Auditability/provenance	**9/10**
Statistical sophistication	**7/10**
Reliability assurance	**6.5/10**
Validity assurance	**6/10**
Ease of first-time use	**6–7/10**
Potential methodological contribution	**8.5/10**

The important point is that **the weaknesses are not primarily weaknesses in the clustering algorithm**.

They are weaknesses in the much more interesting upstream question:

> **How confidently can we regard an LLM-generated criterion as a valid measurement**

> **of a meaningful distinction between texts?**

That is where I think the next phase of the project should concentrate.

My most important recommendation

I would slightly reposition the intellectual centre of gravity of the Lab.

At present it can sound like:

> **"An interpretable alternative to conventional text clustering."**

I think the more interesting proposition is:

> **"A human–LLM method for constructing, testing and using an interpretable typology of texts."**

That puts the **criteria-construction process** at the centre rather than the dendrogram.

The dendrogram then becomes one important consequence of that process.

And I think that is where the real originality lies.

Your own current documentation is already moving in this direction, particularly with the Selection tab, reliability testing, provenance, and the explicit cautions about validity and framing.

One final observation

The fact that this has been developed substantially with an LLM is actually part of what makes the project interesting rather than something that needs to be apologised for. The application is itself an example of the **human–LLM methodological hybridism** that you are investigating.

But that also means I would subject it to a level of methodological testing that goes beyond normal software testing:

Give the same corpus to independent analysts, with several reasonable framings, several LLMs, and some human coding. Then examine where the resulting criteria, matrices and dendrogram structures converge and diverge.

That experiment could tell you considerably more about the Lab's scientific/methodological status than another ten rounds of feature development.

And, importantly, it could produce evidence for a paper arguing that the Lab is not merely *plausible*, but that its classifications have demonstrable **robustness, interpretability and practical utility**.