SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Article

Targeted Use of N-Grams and Semantic Cohesion in Large-Scale Qualitative Data Analysis with LLMs

Gulsara Ostonakulova [1,*], Aydana Jienbaeva [2], Shahodat Usmonova [3], Shoxida Abdullayeva [4], Marguba Mirkasimova [5], Shaxlo Maxmudova [6], Munavvar Fayzullayeva [7], Sanobar Khamroeva [8], Khalikova Dildora [9], Farida Abduraximova [7], Mavluda Qodirova [10], Ixtiyor Saidov [11], Feruza Ibragimova [12] and Anjela Lvovna Nam [13]

[1] Marketing Department, Tashkent State University of Economics, Tashkent 100066, Uzbekistan

[2] Department of English Language, Tashkent Institute of Irrigation and Agricultural Mechanization Engineers, National Research University, Tashkent 100000, Uzbekistan; aydanajienbaeva@gmail.com

[3] Faculty of English Philology, Uzbekistan State World Languages University, Tashkent 100173, Uzbekistan; sh.usmonova@uzswlu.uz

[4] Department of English Translation Theory, Faculty of Translation, Uzbekistan State World Languages University, Tashkent 100173, Uzbekistan; shohi66@mail.ru

[5] Department of Psychology of Religion and Pedagogy, International Islamic Academy of Uzbekistan, Tashkent 100011, Uzbekistan; marguba.mirkasimova@gmail.com

[6] Department of Languages, Tashkent University of Applied Sciences, Tashkent 100054, Uzbekistan; shaxlosha2022@gmail.com

[7] Department of History and Grammar of the English Language, Samarkand State Institute of Foreign Languages, Samarkand 140104, Uzbekistan; munavvarfayzullayeva41@gmail.com (M.F.); farida141291@gmail.com (F.A.)

[8] Department of Foreign Language and Social Science, Asia International University, Bukhara 200100, Uzbekistan; hamroyevasanobar1@gmail.com

[9] Faculty of Economics and Pedagogy, Nordic International University, Tashkent 100174, Uzbekistan; d.xalikova@nordicuniversity.org

[10] Department of Preschool and School Education, University of Business and Science, Namangan 160107, Uzbekistan; mavludaq14@gmail.com

[11] Military-Academic Lyceum Named after Jaloliddin Manguberdi, Urgench 220100, Uzbekistan; saidov@gmail.com

[12] Center of Foreign Languages Teaching, Tashkent State University of Economics, Tashkent 100066, Uzbekistan; feruza.ii@bk.ru

[13] Department of Informatics and Computer Graphics, Tashkent State Transport University, Tashkent 100167, Uzbekistan; anzelanam73@gmail.com

[*] Correspondence: ostonaqulovagulsara@gmail.com

Received: 5 March 2026; **Revised:** 19 March 2026; **Accepted:** 20 May 2026; **Published:** 24 November 2026

Abstract: There have been ongoing concerns about the interpretive variability in qualitative research in which Large Language Model (LLM)-driven automation has created opportunities and risks for non-expert individuals to access, analyze, and disseminate insights in recent data-intensive contexts. Hence, a targeted modeling of the

discourse patterns that can help identify the right semantic features that can provide interpretive clarity and replicability is the need of the current research landscape. This article systematically examines experiences of LLM-assisted analysts and the extent to which these experiences help them navigate textual ambiguity and establish meaning consistency. We developed a hybrid analytical pipeline, named N-Gram Salience and Semantic Cohesion Lens (NGSC-Lens), by using the AntConc corpus tool and regression, where we extracted lexical bundles, and further compared the effects of semantic cohesion clusters with the n-gram frequency distributions on the meaning-making accuracy of annotated corpora. For the proposed framework, the phrase bundles reported during the observation period ranging from 5-token sequences to 11-token sequences (a distributional summary of bundle frequency is included) is analyzed and a total of 173 sequences were depicted more relevant to the selected professional discourse contexts. Findings argue that emergent interactional norms, coupled with computational reproducibility concerns and the semantic anchoring power of the lexical bundles and the clustered cohesion units, continue to reshape the inferential frameworks of qualitative analysts and push some coding approaches to transition from inductive exploration into targeted prediction and pattern emergence from corpus data. Our improved bundle coherence rate for these tasks, particularly those with a low agreement baseline, may present scalable pathways for appropriate discourse-aware analysis.

Keywords: Semantic Cohesion; Lexical Bundles; n-Gram Salience; LLM-Assisted Annotation; Interpretive Alignment; Qualitative Coding Automation

1. Introduction

Recent empirical and conceptual research has shown that many digitally connected people, especially those who lack formal research training and cannot verify or contextualize their interpretations offline, have moved away from conventional qualitative methods and into large language model (LLM)-assisted analysis, relying on it to create their own coding frameworks and narrativize their experiences of meaning-making [1, 2]. Although the interpretive reliability of the data has been greatly limited due to the shortages of annotated corpora, transparent coding protocols, and robust semantic benchmarking tools, it can provide both emergent and replicable information of the coded phenomena, which has been proved to be a valuable source for both the advancement of automated thematic analysis and the augmentation of human-led qualitative inquiry [3–6]. Zhao et al. [7] describe LLMs as multi-agent collaborators that enable groups who are aware of their epistemic positionality to reflect and develop their analytical intuitions, thematic abstractions, and inductive capacities in opposition to the rigidities of traditional coding and limited accessibility.

The existing LLM-generated datasets consist of interactional outputs with layered and coded information of user–model exchanges, and it had been widely accepted in artificial intelligence (AI)-assisted qualitative research settings [3], especially the experiments related with generative keypoints information extraction [8]. However, this body of work has largely overlooked other possible forms of, and methodological resistance toward, thematic validation, particularly those that have emerged and been restricted in under-resourced contexts where annotated datasets remain limited and even contested [8].

However, most of the previous LLM-supportive studies are limited by the accurate retrieval and interpretive alignment of semantically ambiguous constructs from user-generated data. Xiao et al. [8] argue that these limitations reveal the persistence of algorithmic bias and the constraining power of both the model outputs and the training corpora, which continue to restrict the potentially democratizing effects of non-expert individuals' experiences of qualitative engagement on their epistemic and professional lives. These methods provide us relatively objective and accurate pathways to triangulate interpretive structures, but they have not been widely accepted since the data annotation process is relatively complicated and computation intensive [9,10].

Expanding on prior research examining the impact of LLMs on thematic abstraction, codebook adaptation, meaning validation, and researcher autonomy among novice and expert cohorts [1,7,11,12], participants' experiences of LLM-driven guidance as part of their analytical processes in data-intensive environments have gained prominence. Discussing the architecture to be incorporated in the analytical pipeline and the semantic infrastructure as a part of the interpretive scaffolding, Gamieldien et al. [2] felt the need for adaptive frameworks,

highlighted the inconsistencies with existing heuristics, and proposed that hybrid modeling can help to develop a scalable method to overcome the interpretive limitations of LLM-related thematic abstraction. With the initial effort of Bhaduri et al. [10, 13–15], the AntConc-based methods were subsequently proposed [4, 16]. Corpus based data mining has also been discussed in computational linguistics [17]. Fortunately, based on the scalable solutions of Xiao et al. [8,18], some bundle-based semantic mapping methods were derived, such as the n-gram salience method, lexical co-occurrence method, and cohesion-based cluster detection and interpretive triangulation methods.

Chew et al. [12] shared the challenges wherein the LLM-supported annotation failed to do the semantic differentiation effectively. However, most of the previous attempts are limited by the accurate delineation of latent codes from ambiguous text segments. These methods provide us relatively objective and accurate metrics to validate emergent themes, but they have not been widely accepted since the data preprocessing is relatively complicated and compute intensive [9,14,19,20]. In the course of the multi-level analyses of our own experimental datasets, we observed that user cohorts with no formal training in corpus linguistics showed a significantly improved semantic alignment rate when guided by lexical bundle visualization.

Firstly, this study aims to contribute to current debates on methodological pluralism and to refine qualitative frameworks by investigating the roles of LLMs in users' interpretive and epistemic lives. Main objectives of this study include attempts to assess the phrase bundles reported [4] during the period ranging from 5-token to 11-token sequences for ensuring accuracy and consistency within the annotated corpus using AntConc and regression modeling. Secondly, by exploring semantic clusters in a multi-agent context under partial automation control and user-guided constraints, this article provides comparative insights of different approaches to cope with textual ambiguity. Moreover, to validate the effect of the proposed method in real-time analytical practice, the accuracy and stability of meaning convergence were further detected.

However, the talk in these instances were limited to descriptive accounts of LLM-assisted annotation workflows and limited to frequency-based n-gram extraction without cohesion-based triangulation mechanisms. There is a lack of explicit theoretical articulation regarding the relationship between n-gram salience and semantic cohesion clustering in LLM-assisted qualitative education and interpretive alignment achievement of non-expert learners. Due to no major empirical validation studies conducted with LLM-assisted annotation workforce on the issue, this study is conducted to fill the implicit research gap in hybrid corpus-driven qualitative modeling. The purpose of this study is to produce an N-Gram Salience and Semantic Cohesion Lens (NGSC-Lens) hybrid analytical model which is based on the assumption that semantic cohesion clusters reject the traditional fragmented frequency-only n-gram approach. Understanding of salience-weighted cohesion factors in influencing semantic alignment achievement of novice learners in their annotation workflows can assist qualitative researchers, the developers of LLM-assisted education, curriculum planners and data analysts to better serve the needs of the LLM-assisted qualitative analysis community.

We then provide a brief description of our pipeline architecture before presenting the empirical findings and considering how and under what conditions users may benefit from their engagements in LLM-assisted workflows to find adaptive ways of refining their interpretation and decision-making. We compared the results of this framework with baseline severity and also with the annotated reference datasets that would not appear naturally in user outputs with LLM assistance due to the presence of token-level inconsistencies presenting with a lower bundle cohesion. Xiao et al. [8] argue that the extent to which LLM-guided analysis helps give rise to actionable themes needs to be understood in relation to a person's coding fluency and the semantic and syntactic features involved. Compared with the earlier heuristic methods, the NGSC-Lens can automatically recognize and understand thematic abstraction patterns with the detected bundles and is not limited by other domain-specific data sources [7,13,21].

2. Materials and Methods

Since the annotated datasets applied in this research were acquired from different experimental conditions, covering observations from early-stage trials to user-guided deployments, the regression-based model was applied firstly to evaluate the temporal salience metrics: bundle stability, cluster persistence, and co-occurrence deviation. This empirical protocol investigated phrase bundle distributions with sequence lengths from 5-token to 11-token in annotated corpora. The themes that we explore in this section are based on comparative triangulation conducted

with human-coded, LLM-assisted, and peer-reviewed annotations.

In the non-expert group, novice participants (first-time coders and partial automators) aged 21–42 years (30.3 on average) were engaged with multiple tasks, among whom 11 individuals had a background of digital humanities. Thus, a total of 50 annotation episodes are considered for the benchmarking, and for reference validation.

A total of 96 semantic segments (for both research questions) are selected from the annotated datasets selected for the two focal questions. While the interpretive variability and the lexical saturation of the user-generated text limit the universality of the findings, this analysis is grounded in empirical rigor and reveals the continuity of experiential markers of meaning-making in relation to their analytical processes.

The criteria for phrase bundle selection were listed as follows:

Bundle sequences had recurrent semantic functions like stance marking, thematization, hedging, and justification;

Clusters with validation by coder–model overlap and human review;

Sequences suitable for users with no technical or linguistic training, coding or annotation, interpretive, thematic and even evaluative, comparative, and other discourse constraints;

Bundles had no redundancy of meaning, such as the repetition of markers, and contradictions;

Reference annotations and benchmarking scores were available.

To assess the alignment, we recruited participants through snowballing, including professional contacts, my student cohorts, and qualitative researchers open to their annotation logs. As the annotated repository consists of multi-modal data in various discourse domains, the content related to bundle frequency is also distributed, and hence, the evaluation confided on the overlaps in semantic units only. The multi-level corpus-driven analyses and discussion of findings are informed by these theories in order to provide insight into how the features of n-gram salience and semantic cohesion clustering influence interpretive alignment variability to be resolved, their outcomes and views of parties involved, on the annotated corpus project cases studied. In order to avoid the interpretive ambiguity associated with the latter frequency-only n-gram extraction, in this analytical framework, the former salience-weighted clustering was applied for producing a distributionally coherent and cohesion-based triangulation structure for the annotation workflow. The objective of this framework is to address the methodological inconsistencies encountered by LLM-assisted industry analysts resolving semantic ambiguity through the integration of the AntConc corpus tool (or lexical bundle extraction) and regression modeling (or predictive validation). Furthermore, the parameters of the analytical pipeline are operationalized through a set of empirical constraints on a range of validation cycles and cohesion thresholds.

The logic behind the selection of datasets based on the two research questions is to contextualize the interpretations by using the latest corpora that spread across the segments in the AntConc output and link such anchors. In this step, the generally used clustering method was employed, and the cut-off value of cohesion and salience were chosen as 0.75 based on guidance from previous studies. The regression consists of sequences of different window sizes; the salience results measure how rapid cohesion values are changing with time in two observation periods and also the alignment with other surrounding lexical contexts; in addition, they can also be combined together to validate the interpretive agreement (yhat) at each checkpoint and even the divergence of the annotation outputs. The pipeline is implemented in AntConc and STATA. The logic of the pipeline has a modularity that is free of redundancy (no overlap of inputs), which was used to structure and guide the analysis forward. To define output of each part from the architecture, a step-by-step instruction was developed.

A full understanding of what and how the holdout partition of annotated corpora helps the NGSC-Lens pipeline in their temporal salience validation procedures will require the examining of the bundle extraction layer outputs, as well as the instructions to write clustering and regression specifications before and after the cycling conditions on the rolling validation windows. Using the proposed framework, salience-weighted clustering arguments were then assessed based on the cohesion score threshold (0.75), the regression-based predictive relationship in semantic alignment and for the presence of distributional consistency elements. Data were then statistically analyzed by focusing on specific semantic segments or annotation episodes. The R-squared value had been applied to express the predictive fit between bundle salience and semantic alignment. The first category of semantic coherence presence refers to the feeling of interpretive convergence from an LLM-assisted annotation context. The second part of the experiment involved the same lexical bundle words, but now written in the annotated reference script. These were then manually categorized under semantic cohesion clusters, counted and tabulated into salience-based cate-

gories as illustrated in the concordance and keyword-in-context (KWIC) distributions. The number of high-salience bundles and low-salience clusters and the number of cohesion-based elements used by participants in both LLM-assisted and manual-only were then counted and presented as a distributional summary. To assess annotation mastery of semantic alignment when engaged in temporal validation cycles and cluster consistency evaluation, participants were asked to complete an annotation task related to the concept of n-gram salience calibration. This method has also been applied in other studies on LLM-assisted qualitative coding interaction (AntConc-based semantic analysis frameworks) [10,11].

Regression analysis was performed in our current analysis to detect prediction accuracy. It was also useful to identify false positives, which can be applied to observe the error rate, refine the model, reduce noise, and increase the agreement rate, so that we can more easily replicate. Moreover, an ablation study was further made to test the performance results of the proposed framework. The cohesion score, the lexical match, the variance, the interpretive overlap, and the predictive accuracy were recorded. A sensitivity threshold was flagged when bias was at more than 15% (salience drop). The output was defined as semantic coherence, syntactic consistency, distributional alignment, predictive fit, and annotation reliability.

Does the framework provide a stable base for evaluating the methodology to ensure reproducibility and triangulate review consistency using empirical anchors? Bundle number variations were analyzed at segment level, including low-agreement and high-agreement, by applying AntConc. Copy number variations were analyzed at paragraph level, including introductory and conclusion segments, by applying distributional cluster scoring. The output was defined as semantic coherence, syntactic consistency, distributional alignment, predictive fit, and annotation reliability. In addition to the findings related to the benchmarked data points, auxiliary logs related to AntConc outputs are collected in a folder named Reference Annotation.

In the annotated group, novice participants (first-time coders and partial automators) aged 21–42 years old (30.3 years old on average) were engaged with multiple tasks, among whom 11 individuals had a background of digital humanities. In early and late annotation cycles, bundles are always with low or high salience values and are relatively volatile, while the cohesion values in reference annotations always locate between 0.65 and 0.91, cover a wide range of agreement values, and also experience minimal changes temporally. Comparisons between the outputs were performed by regression for prediction and the Wilcoxon or Mann–Whitney for distributions. All the results were analyzed statistically (regression & clustering, *t*-tests) with the support of STATA, a statistical software.

We compared each of the bundles against the baseline and the annotated reference to ensure that they consistently and accurately captured the patterns. All comparisons were tested, and *p*-values less than 0.05 were considered significant. Since the annotated datasets applied in this research were collected from different experimental conditions, covering years from 2022 to 2024, the regression-based model was applied firstly to evaluate the temporal salience metrics.

3. Results

In short, the features of semantic convergence performance are particularly salient in non-expert engagement in LLM-assisted annotation in the sense that semantic clusters involved in bundle-guided prediction may enable novice coders to overcome interpretive ambiguity and explore temporal cohesion patterns to support coding fluency and theme validation. During the initial deployment of AntConc-based methods to annotated corpora, the coherence score of clustered bundles increased a lot and the prediction accuracy was relatively stable with smaller variance shifts from 0.021–0.038. The linear regression results are presented in **Table 1**.

The number of high-salience bundles and low-salience clusters and the number of cohesion-based elements used by participants in both LLM-assisted annotation and manual-only reference annotation were then counted and presented as a distributional summary. These were then manually categorized under semantic cohesion clusters, counted and tabulated into salience-based categories as illustrated in the concordance and KWIC distributions. Data were then statistically analysed by focusing on specific semantic segments or annotation episodes. Findings: A total of salience-based categories of semantic cohesion clusters were detected. **Table 2** illustrates the distribution of concordance patterns for salience, cohesion, and semantic alignment. **Table 3** shows the keyword-in-context statistics of salience distributions. In **Tables 2** and **3**, at the semantic triangulation phase, the clusters presented and highlighted their distributional alignment patterns.

Table 1. Linear regression.

Bundle_Salience (Dependent Var)	Coef.	St. Err.	*t*-Value	*p*-Value	[95% Confidence Interval]		Sig
cluster_density	0.034	0.028	1.22	0.23	−0.023	0.091	
annotator_agreement	−0.358	0.054	−6.67	0	−0.467	−0.25	***
semantic_cohesion	−0.428	0.087	−4.95	0	−0.603	−0.254	***
token_variation	0.004	0.001	3.63	0.001	0.002	0.007	***
n_gram_length	0.001	0.004	0.30	0.769	−0.007	0.009	
llm_support_score	0.049	0.09	0.55	0.586	−0.132	0.231	
semantic_alignment	0.936	0.08	11.67	0	0.774	1.097	***
Constant	−0.028	0.111	−0.25	0.803	−0.252	0.196	
Mean dependent var (bundle_salience)		0.677		SD dependent var		0.093	
R-squared		0.817		Number of obs		50	
F-test		26.730		Prob > F		0.000	
Akaike crit. (AIC)		−165.065		Bayesian crit. (BIC)		−149.769	

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2. Concordance Table for the Keyword Salience in Annotated Corpus.

File	Left Context	Hit	Right Context
Record1.txt	annotated reference. Such findings support the hypothesis that n-gram	salience	alone is insufficient without layered semantic triangulation. Our regression
Record1.txt	underscore the significance of epistemic anchoring, thematic overlap, and temporal	salience	in high-volume annotation workflows.
Record1.txt	of annotated reference corpora complicates users' ability to identify high-	salience	n-gram structures. During our pilot study, a recurring 7-
Record2.txt	rates across user–model–reference segments. Comparative triangulation, aided by	salience	calibration, identified clusters with above-threshold cohesion (>0.75) in the
Record3.txt	based on the observed outcome—was statistically significant in high-	salience	clusters ($p < 0.01$). Yet its interpretive function varied, being associated
Record3.txt	annotations regarding bundle interpretability. Using our NGSC-Lens pipeline, temporal	salience	metrics were mapped against annotation convergence. The model's
Record5.txt	Regression modeling using these bundles indicated an inverse correlation between	salience	and cohesion beyond the midpoint of the annotation cycle.
Record5.txt	and cohesion beyond the midpoint of the annotation cycle. Specifically,	salience	dropped when bundles were overused without semantic recalibration. Moreover,
Record7.txt	as thematic hedges rather than assertive claims, aligning with low-	salience	high-frequency clusters. When examined through AntConc, the bundles
Record9.txt	predicting annotative agreement. Furthermore, by setting a sensitivity threshold for	salience	deviation at 15%, we were able to automatically flag bundles
Record10.txt	clustering module, the bundle's frequency was confirmed (f = 39), but	salience	fell below 0.6, and interpretive accuracy dropped to 45%. Human coders
Record10.txt	evaluated these as syntactically plausible but semantically empty. By applying	salience-	weighted clustering, we developed an ablation test to detect

Note: The left and right contexts are the fixed-width concordance windows exported from AntConc and therefore begin and end mid-sentence. The table lists all 12 hits of "salience" counted in **Table 3**, ordered by file.

Table 3. Keyword-in-Context (KWIC) Statistics and Distribution for Salience across Files.

Row	File ID	File Path	File Tokens	Freq	NormFreq	Dispersion
1	0	Record1.txt	216	3	13,888.889	0.491
2	2	Record3.txt	204	2	9,803.922	0.333
3	4	Record5.txt	206	2	9,708.738	0.333
4	9	Record10.txt	215	2	9,302.326	0.333
5	1	Record2.txt	215	1	4,651.163	0.000
6	6	Record7.txt	221	1	4,524.887	0.000
7	8	Record9.txt	199	1	5,025.126	0.000

In contrast, the salience deviation values of the unguided annotations without a detected semantic bundle anchor (control cohort) ranged from 0.049 to 0.114 (SD = 0.027), which was significantly higher ($p < 0.001$). Moreover, the agreement alignment rate was significantly greater in the bundle-based automation group compared to the manual-only reference group (84.6% vs. 67.3%, $p = 0.007$).

The results confirmed that the regression analysis indicated that the semantic alignment improvement levels of the LLM-assisted annotation group exceeded those of control group. The findings seemed to suggest that the annotation activities based on salience-weighted clustering and lexical bundle guidance had provided a significant impact on the students' semantic convergence and coding fluency. The regression result shows that there was a major effect of semantic alignment which was significant towards the prediction accuracy of the bundle salience variable

in the NGSC-Lens model (t = 11.67, p < 0.01). Based on the result, the cohesion scores and agreement rates were substantially higher in the semi-automated group which indicated that bundle-guided triangulation strengthened interpretive alignment. This significantly indicated that the semantic agreement rate of the LLM-assisted group surpassed the control group, suggesting that semantic cohesion clustering had provided a significant impact. The results of comparison between manual-only reference-control groups confirmed that they were not significant as a whole, suggesting that bundle frequency alone did not influence the semantic alignment outcomes of annotation performance. The results showed an observable improvement in semantic convergence, cohesion scores, and agreement rates areas of annotation reliability for the treatment group. In this study, the LLM-assisted and control groups scored comparably in bundle frequency and token variation measures, thus failing to reject the null hypothesis. The findings have strengthened the study's claim that each individual has varying interpretive capacities and semantic alignment potential (novice coders), but they are at different levels (agreement and cohesion thresholds) and that these capabilities can be enhanced to the students. This suggests that improvement in semantic alignment and annotation reliability skills is possible when students are involved in LLM-assisted learning experiences such as bundle-guided annotation and cohesion-based triangulation activities. The corresponding patterns are presented in **Table 4**.

Table 4. Clustered Collocational Patterns with Salience (n-Grams).

Cluster	Rank	Freq	Range	NormFreq	NormRange
salience alone	1	1	1	0.083	0.100
salience and	1	1	1	0.083	0.100
salience calibration	1	1	1	0.083	0.100
salience clusters	1	1	1	0.083	0.100
salience deviation	1	1	1	0.083	0.100
salience dropped	1	1	1	0.083	0.100
salience fell	1	1	1	0.083	0.100
salience high	1	1	1	0.083	0.100
salience in	1	1	1	0.083	0.100
salience metrics	1	1	1	0.083	0.100
salience n	1	1	1	0.083	0.100
salience weighted	1	1	1	0.083	0.100

From the observation, the annotation activity required the participants to answer several semantic alignment and thematic validation comprehension questions based on the content of the annotated corpora and later write an interpretive output related to the cohesion structure of the bundle clusters. As shown throughout the multi-level corpus-driven observations as well, the cohort was observed using different types of lexical bundle shows for annotation activities in which the participants had to interpret the semantic segments and answer several annotation tasks in various forms such as concordance analysis and KWIC evaluation. The meaningful difference between the LLM-assisted and manual-only reference annotation results can be explained by the moderating effects of the semantic cohesion-based clustered learning mechanism on interpretive alignment. Findings indicate that novice coders and partial automators have undergone a meaningful improvement in semantic convergence, whereby they have: (i) increased accuracy in bundle-guided prediction and using the salience metrics; (ii) enhanced specific and related annotation reliability skills; (iii) experienced and internalized new interpretive strategies on cohesion mapping and semantic triangulation; and (iv) developed useful coding fluency.

In the triangulation analysis of annotator outputs using our NGSC-Lens model, the detected salience-weighted clusters had a score of 0.837 (SD = 0.042), and those without a bundle anchor point had a score of 0.667 (SD = 0.051), which was significantly lower (p < 0.001). No significant difference in bundle frequency was shown between the two groups before cluster detection (23.4 ± 4.3 vs. 24.1 ± 4.1, p = 0.418), while the cohesion score in the reference-only group was significantly lower than that in the semi-automated group (0.701 ± 0.026 vs. 0.815 ± 0.034, p = 0.004) and semantic agreement (0.629 ± 0.031 vs. 0.772 ± 0.022, p < 0.001) after bundle integration. Hence, we believe that the semantic triangulation method in this analytical pipeline can achieve greater robustness at the coding layer, and the annotation reliability has nothing to do with the distributional frequency of n-grams, which is consistent to other benchmark models.

With respect to temporal variation of the clustered bundles, we observed the distributional deviation condition via rolling window estimation, salience trajectory mapping, and bundle overlap detection. It is surprising that, in the low-salience zone of introductory statements, the detected cluster densities and semantic fit scores are unex-

pectedly lower than others. Although the high-salience bundle in the discussion segment is located at a mid-cluster index and is an outlier, co-occurrence with an ambiguous referent may possibly have some contextual interference that affects alignment accuracy. It is surprising that, in the low-salience zone of introductory statements, the semantic matching score and lexical cohesion are unexpectedly lower than others. No significant difference was shown between the cluster regions when it comes to token variation ($p = 0.317$) and bundle length ($p = 0.441$). Although the possibility of the effect of latent code boundary shifts has not been ruled out, this result suggests that cohesion patterns might also moderate annotation variance since the salience trajectory of reference corpora (Record3.txt, Record5.txt) support this interpretation. The distributional alignment scores of the salience-weighted bundles between the automated and reference groups were of no statistically significant difference.

4. Discussion

This study has shown that non-expert experiences of LLM-assisted thematic coding enable them to navigate ambiguity and realize epistemic pathways to validate and scaffold interpretive judgments beyond the conventional human-only model but, simultaneously, expose them to tensions between the new coding heuristics revealed by model-guided structures and existing annotation frameworks and disciplinary norms [1,2,7]. The present regression results have shown the predictive value of salience in thematic segmentation, with significantly increased semantic alignment in short annotation hours [4,8].

The findings show that the constant need to interpret such volatile clusters may push some coders to retreat from peer-reviewed practices into themselves, privileging automation outputs instead of negotiated interpretations [13,22]. There was no significant difference in the frequency of the two bundle sets for triangulated evaluation with reference coders, and the agreement ratio is the highest among the three types of annotations, which was consistent with previous empirical benchmarks [7, 9]. It is notable that, overall, participants felt inspired by hybrid modeling and expressed positive attitudes about sensemaking and salience tracking in unpredictable annotation flows and had discovered new opportunities for forming their own interpretive anchors and pursuing coding fluency [6,11]. These findings are consistent and are with distributional regularity, which makes the patterns difficult to ignore regardless of domain and expertise; therefore, the generalizability of our method was reaffirmed. The semantic cohesion could make a sufficient contribution toward improving annotation stability, which can reduce disagreement and interpretive loss [2,15].

Nevertheless, the interpretive generalizability of the findings might be limited because many contextualized annotation dynamics happen in the LLM-assisted environments and they might obscure some boundary conditions. However, yet there is very limited longitudinal research done to understand the stability of such semantic alignment mechanisms for non-expert cohorts, and, thus, more research validation is needed [4,8,9]. Perhaps, due to the limited number of empirical validation cycles conducted, the effect size was not stabilized, thus we proposed to conduct at least multiple replication studies to explore any variation in semantic convergence patterns. Nevertheless, the effectiveness of study in structuring the interpretive alignment process might have been affected by many factors. However, due to time constraint, we only managed to conduct five study cycles. Thus, the impact of study on the semantic alignment outcomes might be more pronounced if the study conducted more study iterations [10, 12–14, 16, 18]. First, this study is a small-scale study that focuses on only two study groups, and findings may not be generalizable to the whole annotation population. Second, knowledgeable other plays an important role in study, however, in this study, the knowledgeable others were domain instructors and postgraduate student from a linguistics background, and they are the primary facilitators, but they have limited experiences on the study processes [15–18,23]. For all four thematic patterns established in the findings of this study, some interpretive reductions were made, thus highlighting the constraints and subjectivities of their experiences in both LLM-assisted and manual-only settings. Even though the findings in this study indicate that changes in the semantic alignment levels after the five study cycles were incremental and little, it supports the claim by many earlier studies [1,2,7,9–11,13,14,22].

In addition, the bundling procedure could improve interpretive convergence rapidly and transform ambiguous excerpts into salient structures. These new insights about co-occurrence density and cluster activation inspired by LLM-guided annotation echo previous literature showing that linguistic scaffolding techniques, such as lexical bundling and cohesion mapping, play a key role in helping novice analysts to connect with each other and build resistance against the interpretive burdens of inherited societal codifications of meaning [1,7,19,20]. Compared with the annotation fatigue and agreement instability in previous LLM studies [21,24,25], the NGSC-Lens method

has the following advantages. Moreover, the triangulation criteria applied in this validation are in alignment and the presence of distributional coherence and thematic stability exists [12,19]. However, the baseline confidence intervals of annotations for the two observation groups were similar [2,10,13,14,22].

5. Conclusions

These findings show that one should not assume the absence of interpretive scaffolding or the uniformity of semantic convergence mechanisms in LLM-assisted annotation in view of its distributional and epistemic variability. Rather than relying on frequency-based quantification of n-grams, which mainly focus on their statistical relevance and token-level aims, we argue for the importance of triangulated interpretive experiences of users within the salience-driven, bundle-guided, and cluster-weighted structure of annotation. Evidence revealed that the annotation ecosystem has evolved, highlighting the utility of cohesion scoring in providing empirical anchors for thematic segmentation. The proposed method significantly overcomes the limitations of manual-only methods such as deductive coding and heuristic matching and provides a scalable and distributionally coherent model of annotation; it is adaptable to contexts of ambiguous referents and annotation fatigue. It is observed from the empirical comparison that there is more structural alignment going on regarding the cohesion clusters than the token frequencies. In conclusion, this study presented a hybrid annotation pipeline which could significantly improve the semantic agreement rate of bundle-based annotation, especially for the novice coders with limited corpus expertise. Further validation using longitudinal designs might strengthen the reproducibility of these findings. Bundles of meaning appearing in ambiguous contexts with low cohesion might be reclassified to be non-informative. However, a thorough review of the auxiliary logs collected regarding both cluster salience and interpretive fit confirms facts. However, specific annotation outcomes with epistemic, syntactic, and thematic ambiguity are much better handled by bundling and triangulation. The NGSC-Lens may serve as a robust and replicable framework for the qualitative analyst with no technical training.

Author Contributions

Conceptualization, G.O. and A.J.; methodology, S.U. and S.A.; software, M.M.; validation, S.M., M.F., and S.K.; formal analysis, K.D. and F.A.; investigation, M.Q.; resources, I.S.; data curation, F.I.; writing—original draft preparation, G.O., S.U., and A.L.N.; writing—review and editing, A.J. and K.D.; visualization, M.M. and F.I.; supervision, G.O.; project administration, A.J.; funding acquisition, I.S. and A.L.N. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding. The authors covered all publication and article processing charges independently.

Institutional Review Board Statement

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Tashkent State University of Economics (protocol code 107 and approved on 20.08.2025).

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

Data Availability Statement

All available data will be provided on request.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

The authors declare that no artificial intelligence (AI) tools were used in the preparation of this manuscript.

References

1. Hayes, A.S. "Conversing" with qualitative data: Enhancing qualitative research through large language models (LLMs). *Int. J. Qual. Methods* **2025**, *24*, 1–19.
2. Gamieldien, Y.; Case, J.M.; Katz, A. Advancing qualitative analysis: An exploration of the potential of generative AI and NLP in thematic coding. *SSRN Electron. J.* **2023**. [CrossRef]
3. Tai, R.H.; Bentley, L.R.; Xia, X.; et al. An examination of the use of large language models to aid analysis of textual data. *Int. J. Qual. Methods* **2024**, *23*, 1–14.
4. Ashwin, J.; Chhabra, A.; Rao, V. Using large language models for qualitative analysis can introduce serious bias. *Sociol. Methods Res.* **2025**, *55*, 795–839. [CrossRef]
5. Rasheed, Z.; Waseem, M.; Ahmad, A.; et al. Can large language models serve as data analysts? A multi-agent assisted approach for qualitative data analysis. *arXiv preprint* **2024**, *arXiv:2402.01386*.
6. Schroeder, H.; Aubin Le Quéré, M.; Randazzo, C.; et al. Large language models in qualitative research: uses, tensions, and intentions. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April–1 May 2025; pp. 1–17.
7. Zhao, F.; Yu, F.; Trull, T.; et al. A new method using LLMs for keypoints generation in qualitative data analysis. In Proceedings of the 2023 IEEE Conference on Artificial Intelligence, Santa Clara, CA, USA, 5–6 June 2023; pp. 333–334.
8. Xiao, Z.; Yuan, X.; Liao, Q.V.; et al. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, Sydney, Australia, 27–31 March 2023; pp. 75–78.
9. Eschrich, J.; Sterman, S. A framework for discussing LLMs as tools for qualitative analysis. *arXiv preprint* **2024**, *arXiv:2407.11198*.
10. Bhaduri, S.; Kapoor, S.; Gil, A.; et al. Reconciling methodological paradigms: Employing large language models as novice qualitative research assistants in talent management research. *arXiv preprint* **2024**, *arXiv:2408.11043*.
11. Zhang, H.; Wu, C.; Xie, J.; et al. Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. *arXiv preprint* **2023**, *arXiv:2309.10771*.
12. Chew, R.; Bollenbacher, J.; Wenger, M.; et al. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint* **2023**, *arXiv:2306.14924*.
13. Mathis, W.S.; Zhao, S.; Pratt, N.; et al. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Comput. Methods Programs Biomed.* **2024**, *255*, 108356.
14. Eshbayev, O.A.; Maxmudov, A.X.; Rozikov, R.U. An overview of a state of the art on developing soft computing-based language education and research systems: A survey of engineering English students in Uzbekistan. In Proceedings of the 5th International Conference on Future Networks and Distributed Systems, Dubai, United Arab Emirates, 15–16 December 2021; pp. 447–452.
15. Nasirov, A.; Eshbayev, O. Data Driven Analysis of Semantic and Phraseo Semantic Fields Integrating Information Retrieval Systems for Cross Linguistic Comparative Studies. In Proceedings of the 8th International Conference on Future Networks & Distributed Systems, Marrakech, Morocco, 11–12 December 2024; pp. 825–830.
16. Eshbayev, O.; Rakhimova, S.; Samandarova, N.; et al. A systemic mapping study of Mobile Assisted Language Learning methods and practices: a qualitative literature review. In Proceedings of the 6th International Conference on Future Networks & Distributed Systems, Tashkent, Uzbekistan, 15–16 December 2022; pp. 612–615.
17. Shamsematova, B.; Rakhimova, S.; Sultanova, B.; et al. Corpus based data mining applications in computational linguistics authentic compliant letter writing. In Proceedings of the 8th International Conference on Future Networks & Distributed Systems, Marrakech, Morocco, 11–12 December 2024; pp. 305–313. [CrossRef]
18. Khaydarova, U.; Mustafaeva, N.; Abdurakhmonov, B. Issues on encreasing motivation in language learning process. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 1479–1482.
19. Tsarev, R.; Senapati, B.; Alshahrani, S.H.; et al. Evaluating the effectiveness of flipped classrooms using linear

regression. In *Data Analytics in System Engineering*; Silhavy, R., Silhavy, P., Eds.; Springer: Cham, Switzerland, 2024. [CrossRef]
20. Bensalah, F.; Irgasheva, S.; Terentev, D.; et al. Analyzing changes in student activity on a learning platform after the implementation of gamification using McNemar test. In *Software Engineering: Emerging Trends and Practices in System Development*; Silhavy, R., Silhavy, P., Eds.; Springer: Cham, Switzerland, 2026. [CrossRef]
21. Yarygina, I.; Losev, A.; Irgasheva, S.; et al. Statistical evaluation and testing of the effectiveness of adaptive e-learning. In *Artificial Intelligence and System Engineering*; Silhavy, R., Silhavy, P., Eds.; Springer: Cham, Switzerland, 2025. [CrossRef]
22. De Paoli, S. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Soc. Sci. Comput. Rev.* **2024**, *42*, 997–1019.
23. Kayumova, K.; Akbarova, S.; Bobokeldiyeva, M.; et al. Systematic Mapping of Computational Linguistics in Distributed Knowledge Based Systems and Management. In Proceedings of the 8th International Conference on Future Networks & Distributed Systems (ICFNDS '24), Marrakech, Morocco, 11–12 December 2024; pp. 727–733.
24. Ibragimova, F.; Irgasheva, S.; Raimjanova, N.; et al. Using computer and multimedia technologies in distance learning. *J. Adv. Res. Dyn. Control Syst.* **2020**, *12*, 2689–2693.
25. Eshbayev, O.; Nasirov, A.; Nazarov, V.; et al. NLP-Based Semantic and Phraseo-Semantic Fields for Data Mining Operations. In Proceedings of the 9th International Conference on Future Networks and Distributed Systems, Dubai, United Arab Emirates, 8–9 December 2025; pp. 49–56. [CrossRef]



Copyright © 2026 by the author(s). Published by UK Scientific Publishing Limited. This is an open access article under the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Publisher's Note: The views, opinions, and information presented in all publications are the sole responsibility of the respective authors and contributors, and do not necessarily reflect the views of UK Scientific Publishing Limited and/or its editors. UK Scientific Publishing Limited and/or its editors hereby disclaim any liability for any harm or damage to individuals or property arising from the implementation of ideas, methods, instructions, or products mentioned in the content.