Diagnosing the Cultural Gap in Machine Translation of Classical Chinese: A Computational Analysis of *The Analects*

Menghan Dong, YouzhiWu, SiqiWang, Zhen Wu

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

Diagnosing the Cultural Gap in Machine Translation of Classical Chinese: A Computational Analysis of *The Analects*

Menghan Dong* **and Youzhi Wu and Siqi Wang and Zhen Wu**
The Hong Kong Polytechnic University
24072087g@connect.polyu.hk,
24067312g@connect.polyu.hk,
24104191g@connect.polyu.hk,
24138226g@connect.polyu.hk

Abstract

This paper presents a computational diagnostic of how current AI systems translate classical Chinese, using *The Analects* as a case study. We formalize a multi-view evaluation framework that integrates lexical overlap (Jaccard similarity), distributional semantics (cosine similarity), keyword salience (TF–IDF), topic structure (LDA), semantic networks, and sentiment analysis. Applying the framework to a representative human translation (Legge) and an AI translation (ChatGPT), we find high distributional consistency (cosine ≈0.91) but low lexical overlap (Jaccard ≈0.33), fragmented thematic structure in AI outputs (7 topics vs. 2 in the human translation), and attenuated affective intensity. A qualitative error taxonomy further reveals challenges unique to classical Chinese: homophonic loan characters, polysemy and sense selection, ancient–modern semantic drift, ellipsis/inversion, and value-laden terms. Our study complements standard MT evaluation by moving beyond surface-form metrics (e.g., n-gram overlap or learned estimators) to operationalize cultural and thematic fidelity for classical Chinese. We discuss how this multi-view diagnostic can guide sense-aware MT evaluation and modeling for low-resource, pre-modern language varieties.

1 Introduction

In recent years, artificial intelligence (AI) has improved translation, including literary translation, by increasing speed and efficiency and raising accuracy in specific settings. This progress has broadened access to literary works and provided technical support for translators, while also enabling wider access to Chinese materials for non-Chinese readers. At the same time, evaluating translation quality for pre-modern language varieties such as classical Chinese remains difficult: surface n-gram metrics and even learned estimators often miss sense selection, value-laden terminology, and culturally anchored themes. To examine these issues concretely, we focus on the English translation of *The Analects* as a case study. Our goal is to assess how effectively AI captures thematic depth and cultural subtleties in a classical text. We hypothesize that AI translation can capture major themes in *The Analects*, and we test this through a comparative analysis of AI generated translations and those produced by human scholars, with attention to the accuracy and depth of thematic and ethical representation. Our objective is to propose and validate a model-independent diagnostic framework for assessing systems' understanding of Classical Chinese, rather than to rank the translation performance of specific models. Accordingly, we adopt a widely accessible general-purpose large language model (ChatGPT) as the reference baseline and compare its output with a human translation. Domain-specific models for Classical Chinese (e.g., "XunziLLM") may involve uncertainties in version availability, interface stability, and licensing, making it difficult in the short term to meet the level of reproducibility required for our experiments. To preserve comparability under identical preprocessing and evaluation pipelines, we therefore prioritize publicly accessible models with stable access. We use James Legge's translation as the human reference, and we generated AI translations with ChatGPT and Tongyi Qianwen; after comparing the two, we retain ChatGPT's output for detailed analysis. Legge's translation adopts smaller units that follow original order and is widely treated as a standard version, and in our study it serves as a culturally grounded reference rather than a sen-

[*] Corresponding author

tence level gold standard. Our research questions are as follows: RQ1 (Semantic consistency): How closely-aligned are the AI and human translations at the distributional semantic level? RQ2 (Lexicon and themes): Do the AI translations exhibit systematic deviations in lexicalization and thematic structure (e.g., term generalization, topic fragmentation)? RQ3 (Affect and value terms): Do the AI translations attenuate affective intensity or mishandle value-laden terminology central to Confucian ethics?

2 Literature Review

Artificial intelligence (AI) translation tools have advanced rapidly and expanded assistive use in literary translation, yet they do not replace human expertise for culturally dense texts (Massion, 2017). From an NLP perspective, evaluation remains difficult: standard automatic metrics (BLEU, TER, chrF; learned metrics such as BERTScore, COMET) emphasize surface overlap or sentence level adequacy/fluency but under represent sense selection, value-laden terms, allusions, and culturally anchored themes typical of classical texts. Empirical comparisons reflect this gap. Ding (2024) contrasts human and AI renderings of Li Bai's "Farewell to a Friend," noting human advantages in stylistic nuance and lexical choice, with AI useful for research/learning but uneven on poetic effects. Zaid and Bennoudi (2023) compare human, ChatGPT, and Google Translate on Arabic religious texts, finding automatic systems fairly accurate yet weaker on depth, cultural relevance, and nuanced understanding—dimensions only partially reflected by automatic scores. Al Sawi and Allam (2024) analyze Arabic subtitles for Birdman and show that humans better handle allusions and cultural cues, while AI exhibits limitations on culturally complex content. In NLP evaluation, this motivates complements to surface metrics—distributional similarity, topic modeling, co-occurrence networks, and sentiment/affect profiling—which better probe lexical, thematic, and affective fidelity in classical material. Our study follows this line, using *The Analects* as a testbed to examine where AI aligns with a human reference and where it drifts in lexicon, themes, and affect.

3 Data and Methodology

3.1 Data Sources

We selected James Legge's public domain translation as the human reference and used Tongyi Qianwen and ChatGPT to translate *The Analects*. After evaluating and comparing the two AI translations, we retained ChatGPT's output as the representative machine translation for this study. In the observations, Qwen frequently (i) mixes bilingual translation (e.g., pinyin with parenthetical glosses explanation), (ii) introduces interpretive or modernized paraphrasing, and (iii) alternates near-synonyms or explanatory term pairs for single concepts, which will interfere the accuracy and affect the stability of the following measurement.

3.2 Collection Methods

We extracted Legge's translation from Ctext and submitted the original classical Chinese text of *The Analects* to the AI systems on a chapter-aligned basis to obtain machine translations. All translations were saved as UTF 8 plain text (.txt) files.

3.3 Processing Techniques

We used R and Python to preprocess the .txt files. Steps included Unicode normalization, normalization of quotation marks and hyphens, lowercasing, punctuation removal (except apostrophes), whitespace cleanup (removing redundant spaces), and regex-based English tokenization (RegexpTokenizer). Stopwords were removed for specific analyses; for LDA (and overlap) we used an expanded list including domain-dominant terms (e.g., confucius, master, gentleman, said, zi), while for TF–IDF we used only general English stopwords. Lemmatization was applied for LDA.

3.4 Data Analysis

Using Python and R, we conducted the following analyses: word frequency, word cloud, cosine similarity, Jaccard similarity, TF–IDF, LDA topic modeling, semantic co-occurrence network analysis, and sentiment/affect analysis. Section 4 details each component (4.1–4.7).

3.5 Task Formulation and Metrics

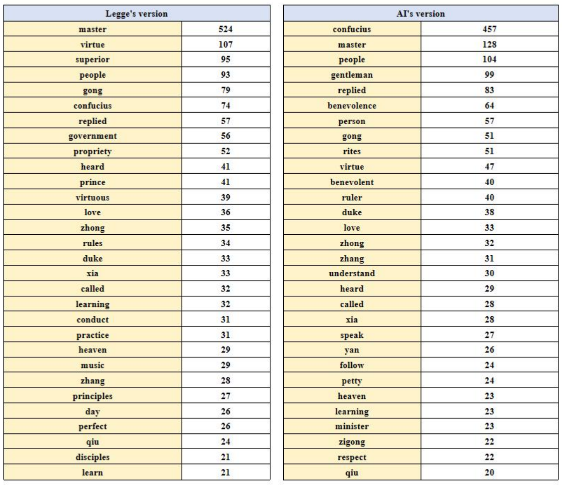
Task. Given a classical Chinese source segmented by canonical chapters, a human translation (H) and an AI translation (M) in the same target language, we quantify the cultural/thematic fidelity of M relative to H across complementary views.

Metrics. Lexical overlap: Jaccard similarity over content word vocabularies after stopword removal (optional lemmatization). Distributional semantics: cosine similarity between TF–IDF vectors at chapter and document levels. Keyword salience: compare TF–IDF top-100 terms and weight concentrations across H and M. Thematic structure: LDA topic modeling; select optimal K via coherence and compare K(H) vs. K(M) and topic separability. Semantic networks: co-occurrence graphs; compare core node centrality and community structure (e.g., modularity). Affective profile: lexicon based polarity/intensity distributions at the chapter level. Qualitative checks: classical specific phenomena (ellipsis, inversion, ancient–modern semantic drift, polysemy, value-laden terms). Stopwords are removed for overlap and LDA with an expanded domain-specific list, while TF–IDF uses only general stopwords; lemmatization is applied for LDA (optional elsewhere). Random seeds are fixed for vectorization and LDA, and chapter level metrics are aggregated to document level summaries.

4 Multi-View Diagnostic Analysis

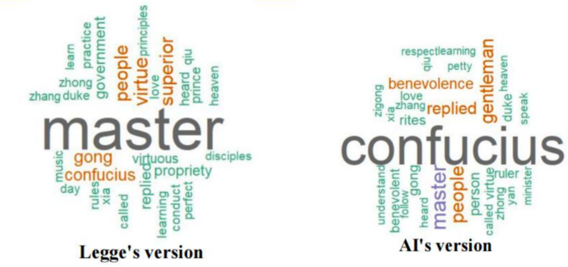
4.1 Word Frequency and Word Cloud

By conducting a word frequency computing and word cloud analysis, we can observe the similarities and differences in lexical usage between the two translations. The statistical results show that there are certain similarities in the distribution of high-frequency words between these two versions. For example, the words “Confucius” and “Master” both refer to Confucius, and “virtue” refers to personal morality. “Superior” and “gentleman” refer to the concept of “junzi,” “government” and “ruler” refer to the sovereign, and these words all appear frequently. This shows that AI translation can capture the core concepts and main ideas of *The Analects*. However, further analysis reveals differences in translation tendencies and vocabulary selection between the two. In the human translation, high-frequency words are closely related to the themes of the superior man and social governance, such as “propriety” and “virtuous.” These words reflect the high emphasis on moral norms and ideal personality in *The Analects* and show the translator’s profound understanding of the original text. In contrast, the high-frequency words in AI translation incorporate more modernized words, such as “gentleman,” “benevolence,” and “rites,” which are highly abstract and general words that are easy to understand. These words are related to the core concepts of the text, but the choice of words is more inclined toward general expressions and fails to fully reflect the cultural context and philosophical depth of *The Analects*.



Legge's version		AI's version	
master	524	confucius	457
virtue	107	master	128
superior	95	people	104
people	93	gentleman	99
gong	79	replied	83
confucius	74	benevolence	64
replied	57	person	57
government	56	gong	51
propriety	52	rites	51
heard	41	virtue	47
prince	41	benevolent	40
virtuous	39	ruler	40
love	36	duke	38
zhong	35	love	33
rules	34	zhong	32
duke	33	zhang	31
xia	33	understand	30
called	32	heard	29
learning	32	called	28
conduct	31	xia	28
practice	31	speak	27
heaven	29	yan	26
music	29	follow	24
zhang	28	petty	24
principles	27	heaven	23
day	26	learning	23
perfect	26	minister	23
qiu	24	zigong	22
disciples	21	respect	22
learn	21	qiu	20

Figure 1: Top-word Frequency Comparison



Legge's version

AI's version

Figure 2: Word Cloud

4.2 Cosine Similarity

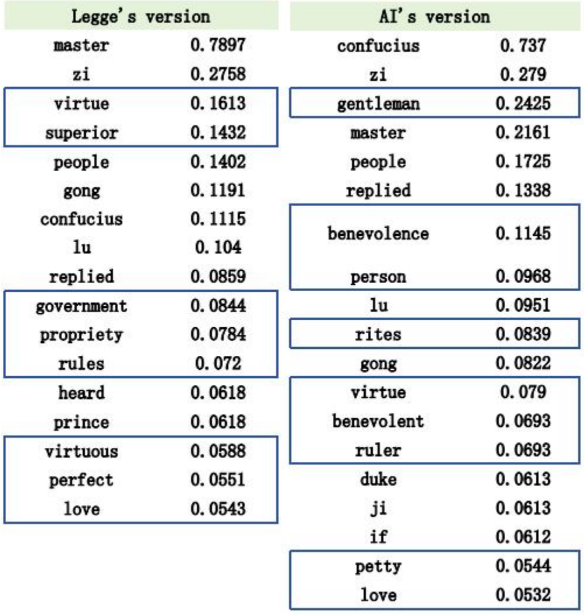
Next, we conducted semantic consistency analysis to statistically measure the degree of similarity between the two translations. We used two indicators: cosine similarity and Jaccard similarity. Cosine similarity is a similarity measurement tool for vectorized text that evaluates the consistency of text in its overall semantic structure. In our study, the cosine similarity between AI and human translation was 0.906, indicating high consistency in document-level lexical distribution and topical signal (TF–IDF vectors), rather than sentence-level structure. This means that AI translation can translate accurately in terms of sentence meaning, semantic logic, and content expression.

4.3 Jaccard Similarity

Jaccard similarity places more emphasis on the degree of overlap in vocabulary selection. However, the Jaccard similarity was only 0.325, demonstrating significant differences in specific vocabulary and phrases between the two translations. It was observed that human translation tends to choose words that fit the context of ancient Chinese and pays more attention to the cultural connotations, such as explaining and interpreting the meaning of a core concept rather than summarizing it directly. For instance, “junzi” is rendered as “a man of complete virtue.” AI translation tends to choose direct or generalized expressions, especially using modern or more universally applicable words to express a core concept with a single word. For instance, “junzi” is rendered as “gentleman.” This is because AI translation lacks a deep understanding of the historical context and cultural background of the text, and its word choice tends to be more uniform and general. Human translation, on the other hand, can combine the context of the text and use flexible word choice and more detailed explanations to more accurately convey the core ideas of the text. Through a combined analysis of cosine similarity and Jaccard similarity, it can be seen that AI translation has good semantic consistency and can convey the logical structure and main content of *The Analects* to a large extent. However, its limitations in vocabulary selection are obvious, resulting in low word overlap.

4.4 TF–IDF Analysis

Through TF–IDF analysis, the differences in keyword extraction and the distribution of vocabulary weight between the two translations can be clearly identified. We compute document-level TF–IDF with scikit-learn TfidfVectorizer on a two-document corpus (human vs. AI), after lowercasing, punctuation and digit removal, and stopword filtering; weights are L2-normalized per document, and we report top-k terms (k = 100). TF–IDF is a method for reflecting the relative importance of a word by calculating its frequency of occurrence in a single document and its distribution in the entire corpus. The high TF–IDF words in the human translation include “master” (0.7897), “virtue” (0.1613), and “superior” (0.1432), which reflect the core ethical and philosophical concepts of Confucius in *The Analects*, such as the concept of the “junzi” and “morality.” The high TF–IDF words in the AI translation include “Confucius” (0.737), “gentleman” (0.2425), and “benevolence” (0.1145). It can be seen that the key words in both translations are highly consistent in content theme, reflecting a focus on the core concepts of *The Analects*, but with different specific choices of words. Focusing on the weights, although the AI translation can capture the keywords and translate them in a way that is close to the human translation, its allocation of weight to these keywords is more balanced and even, lacking prominence and emphasis, indicating that it still has a problem in theme identification.



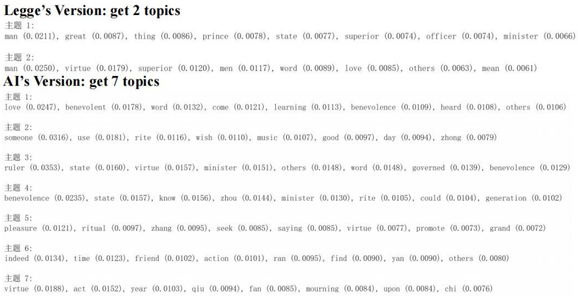
Legge's version	
master	0.7897
zi	0.2758
virtue	0.1613
superior	0.1432
people	0.1402
gong	0.1191
confucius	0.1115
lu	0.104
replied	0.0859
government	0.0844
propriety	0.0784
rules	0.072
heard	0.0618
prince	0.0618
virtuous	0.0588
perfect	0.0551
love	0.0543

AI's version	
confucius	0.737
zi	0.279
gentleman	0.2425
master	0.2161
people	0.1725
replied	0.1338
benevolence	0.1145
person	0.0968
lu	0.0951
rites	0.0839
gong	0.0822
virtue	0.079
benevolent	0.0693
ruler	0.0693
duke	0.0613
ji	0.0613
if	0.0612
petty	0.0544
love	0.0532

Figure 3: TF-IDF

4.5 LDA Topic Modeling

LDA topic modeling reveals the difference in theme extraction. For LDA, we segment the translation into sentence-like units by punctuation and retain segments longer than 50 characters, yielding 907 document units (7,395 tokens) for the AI translation. We select the number of topics via c_v coherence under a lightweight screening configuration (fixed seed), and then retrain the final model with a stronger configuration. Under this setting, the selected number of topics for the human translation is two, and the themes are concentrated on two core areas of “social governance” (state, officer, minister) and “personal morality” (virtue, superior, love). The boundaries between these two themes are very clear, reflecting the translator’s deep understanding of the main themes of *The Analects*. On the other hand, the themes extracted by AI translation are more dispersed, with a total of seven themes identified. Some of the topic words do not have clear practical meanings, and there are also problems of semantic overlap or blurred boundaries between the themes. This phenomenon of semantic and theme fragmentation indicates that AI translation has weak theme coherence in extracting deep-level semantics. Therefore, AI translation lacks proficiency in grasping the overall semantic meaning, has limited understanding of the deep cultural connotations, and has a weaker ability to maintain semantic coherence when dealing with high-difficulty texts such as ancient Chinese literature.



Legge's Version: get 2 topics

主题 1:
man (0.0211)， great (0.0087)， thing (0.0086)， prince (0.0078)， state (0.0077)， superior (0.0074)， officer (0.0074)， minister (0.0066)

主题 2:
man (0.0250)， virtue (0.0179)， superior (0.0120)， men (0.0117)， word (0.0089)， love (0.0085)， others (0.0063)， mean (0.0061)

AI's Version: get 7 topics

主题 1:
love (0.0247)， benevolent (0.0178)， word (0.0132)， come (0.0121)， learning (0.0113)， benevolence (0.0109)， heard (0.0108)， others (0.0106)

主题 2:
someone (0.0316)， use (0.0181)， rite (0.0116)， wish (0.0110)， music (0.0107)， good (0.0097)， day (0.0094)， zhong (0.0079)

主题 3:
ruler (0.0353)， state (0.0160)， virtue (0.0157)， minister (0.0151)， others (0.0148)， word (0.0148)， governed (0.0139)， benevolence (0.0129)

主题 4:
benevolence (0.0235)， state (0.0157)， know (0.0156)， zhou (0.0144)， minister (0.0130)， rite (0.0105)， could (0.0104)， generation (0.0102)

主题 5:
pleasure (0.0121)， ritual (0.0097)， zhang (0.0095)， seek (0.0085)， saying (0.0085)， virtue (0.0077)， promote (0.0073)， grand (0.0072)

主题 6:
indeed (0.0134)， time (0.0123)， friend (0.0102)， action (0.0101)， ran (0.0095)， find (0.0090)， yan (0.0090)， others (0.0080)

主题 7:
virtue (0.0188)， act (0.0152)， year (0.0103)， qiu (0.0094)， fan (0.0085)， mourning (0.0084)， upon (0.0084)， chi (0.0076)

Figure 4: LDA Topic Modeling

4.6 Semantic Networks

Semantic network analysis visualizes conceptual co-occurrence and helps examine community structure and the separation of conceptual clusters. Core concepts (such as "truth," "benevolence," "wisdom") play a central role in both translations, indicating that both of them can capture important semantic nodes in the text. The semantic network structure of the human-translated text is clearer, with nodes distributed more evenly, and core concepts such as "truthfulness" and "benevolence" being more independent from other nodes. The connections between positive and negative words are more distinct, presenting a clearer semantic logic that reflects the accurate grasp of ethical reasoning in *The Analects*. In contrast, the AI network appears denser with lower apparent modular separation among conceptual clusters, suggesting weaker separation of thematic communities. The boundaries between positive and negative emotional words are blurred, and the connections between core concepts are high in density but lack logical coherence. This demonstrates that AI translation tends to handle affect in a generalized manner and tends to summarize relationships between words, failing to fully reflect the clarity of ethical oppositions in *The Analects*.

4.7 Sentiment Analysis

Sentiment analysis results show that the indices are close between these two versions, but the values for AI are generally lower than the human version. AI translation exhibits a general phenomenon of "intensity decay," while maintaining the overall emotional distribution pattern. This "intensity decay" may stem from AI translation's tendency to opt for "safer" lexical and semantic choices. The intensity may be gradually reduced due to probability smoothing, which shows more "conservatism" and "uncertainty" in the translation. While this conservatism ensures stability and reliability, it also weakens the expression of emotional intensity and ethical opposition.

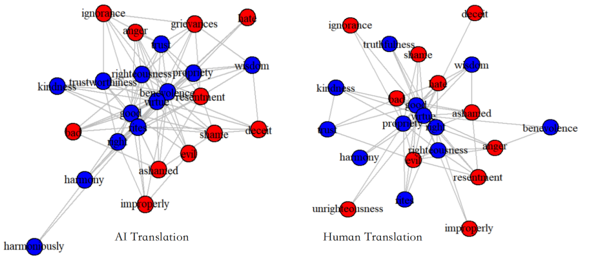


Figure 5: Semantic Network

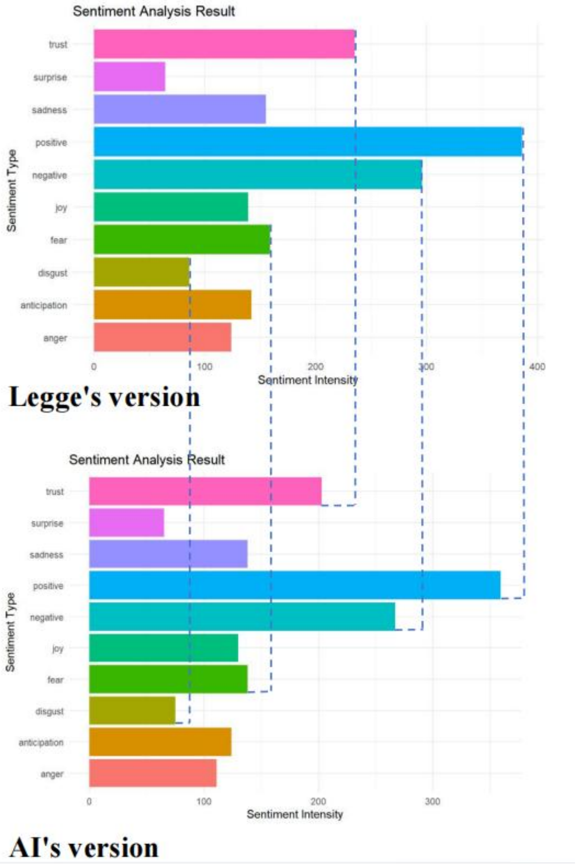


Figure 6: Sentiment Analysis

5 Targeted Case Analyses on Classical Chinese Phenomena

Case analysis primarily explores the potential details and problems that AI translation may encounter when processing *The Analects* text, specifically including loan characters (substitute characters with similar pronunciation), polysemy, special sentence structures, and value embodiment. This part is not suitable for quantitative research, so we adopt qualitative analysis. Loan Characters (substitute characters with similar pronunciation): When dealing with homophones, AI translation tends to give literal translations without fully understanding the context of the text. For example, "樊迟问知" (Fan Chi asks about what is wisdom), "知" (knowledge) is a homophone of "智" (wisdom). However, it is translated as "knowledge" by AI translation, while human translation would more accurately choose "wisdom" to convey the deep meaning in the ethical context. Characters with Different Meanings in Ancient and Modern Chinese For the translation of words with different meanings in ancient and modern times, AI translation performs well. For example, in the sentence "人皆有兄弟，我独亡。" (All men have brothers, but I have none), the word "亡" is translated accurately by AI as "have none," which means "without." Polysemy: One word with multiple meanings in classical Chinese is a difficult aspect in translation, such as "dao," which can represent "road," "principle," or "method" in different contexts. For example, in the sentence "君子所贵乎道者三," the human translator handled "dao" as "principles of conduct," accurately conveying its meaning, indicating the principles that the morally exemplary person should adhere to and keep doing, while the AI translation tends to translate it directly as "way," which cannot fully convey its deep meaning in the specific context. Special Sentence Structure: There are often elliptical sentences and inverted sentences in *The Analects*, which are not easy to understand or translate. For example, "君子上达，小人下达。" Human translators can complete the implicit elements to fully present the logical relationship of the sentence, while AI translation directly renders it as "Gentlemen reach high positions, while petty people reach low positions," translating it as a difference in social status rather than moral levels, thus failing to accurately convey the core ideology of Confucius. Values: The core theme of *The Analects* lies in the conveying of ethics and values, such as the phrase "有教无类," which conveys the idea that education can narrow the moral gap between different people. AI translation translates it as "Education should have no class distinction," which is a literal translation, assuming that people from any class—whether high or low—can receive equal education, which does not take into account the actual historical background at that time, when only nobles could receive a good education.

6 Limitations

Single reference translation: The study mainly relies on James Legge's English version. One reference can't cover the range of translator styles and term systems, which may shift keyword weights, topic distributions, and affect strength, and in turn change the robustness of the findings. Metrics focuses mainly on lexical and phrase distributions: The current framework doesn't fully capture sense allusion, rhetoric, and discourse structure. Future work will add consistency checks based on sentence and paragraph embeddings and run small expert evaluations with annotated samples. Corpus and genre scope: The analysis uses only the Analects, which is aphoristic in form. Other classics and genres may change topic granularity, discourse organization, and terminology expression. It can be extended across texts and genres to other ancient classics.

7 Conclusion

Our analyses indicate strong alignment in the overall topical signal between the AI translation and the human reference (document-level TF–IDF cosine = 0.906), yet systematic divergences in lexicalization and thematic organization remain: lexical overlap is limited (Jaccard = 0.325), keyword salience is flatter in the AI output, LDA reveals fragmented and less separable topics (H: K = 2 with coherent "social governance" and "personal morality"; M: K = 7 with overlap), semantic networks are denser and less modular, and affective intensity is attenuated relative to the human translation. Taken together, these findings suggest that contemporary systems can capture major themes of *The Analects* while diluting cultural anchoring and ethical nuance, addressing our research questions by demonstrating high distributional alignment alongside inaccuracies in concept realization, thematic cohesion, and affect. Methodologically, the study contributes a reproducible evaluation pipeline that integrates distributional similarity, lexical overlap and keyness, topic modeling, semantic networks, sentiment profiling, and targeted diagnostics tailored to classical Chinese, underscoring that surface or distributional scores alone can overestimate adequacy for culturally dense material. Limitations include reliance on a single primary human reference, one retained AI system, English-side analyses, and lexicon-based sentiment. Future work will incorporate learned metrics, sense- and allusion-aware representations, graph-based coherence measures, expert human evaluation with multi-reference targets, and extensions to other pre-modern genres, with code and processed data released for reproducibility.

References

I. Al Sawi and R. Allam. 2024. Exploring challenges in audiovisual translation: A comparative analysis of human- and ai-generated arabic subtitles in Birdman. *PLOS ONE*, 19(10):e0311020.

M. Ding. 2024. Comparative analysis of classical chinese poetry translation using artificial intelligence: A case study of different english versions of li bai's "farewell to a friend". *Education Journal*.

James Legge. 1861. *The Chinese Classics: Confucian Analects*. Trübner & Co.

F. Massion. 2017. Artificial intelligence, smart assistants and the role of language professionals. *Lebende Sprachen*, 62(2).

D. Wang. 2008. Comparative study of english translations of the "analects". Master's thesis, Shandong University, Jinan, China.

A. Zaid and H. Bennoudi. 2023. Ai vs. human translators: Navigating the complex world of religious texts and cultural sensitivity. *International Journal of Linguistics, Literature and Translation*, 6(11):173–182.