# The Collision between AI and Old Uighur Manuscripts: Opportunities and Challenges in the Dunhuang Academy Collection

**Abudouriyimu Rousitaimujiang***
*Dunhuang Textual Research Institute, Dunhuang Academy, Lanzhou, China*
**Corresponding Author*

**Abstract: The Old Uyghur manuscripts collected by Dunhuang Academy are among China's most significant such relics, divided into two main collections: early collected manuscripts and those unearthed from the Northern Area of Mogao Grottoes. The materials under discussion here include Buddhist scriptures, literary works and socio-economic documents, and they are considered to be of the utmost importance for the study of the history and culture of ancient ethnic groups in Northwest China. The subject has been long constrained by fragmentation, with many fragments remaining unclassified and unpublished due to difficulties in deciphering and interpreting the extant material. This is compounded by the paucity of scholars specialising in the field, resulting in a lack of expertise and understanding. The integration of artificial intelligence with image recognition, big data analysis and intelligent collation offers a novel technical approach. The present paper methodically organizes the sources and research status of these manuscripts, with a focus on the opportunities and challenges of integrating AI with manuscript research. The utilisation of artificial intelligence has been demonstrated to yield significant benefits in the domains of fragment assembly, character recognition, digitisation, and comparison. The application of deep learning-based multi-modal feature extraction has been demonstrated to accelerate fragment collation, with convolutional neural networks achieving over 96% accuracy in the recognition of ancient Old Turkic script. The utilisation of intelligent databases and big data comparison has been demonstrated to enhance efficiency and reveal hitherto overlooked textual connections. Nevertheless, the process of deep integration continues to encounter significant challenges. The extinct language known as Old Uyghur is characterised by damaged documents, inconsistent norms and scarce training data. It is evident that artificial intelligence is not yet capable of comprehending religious terminology, cultural nuances, and academic logic with the same depth and breadth as human beings. The proposed methodology involves the construction of high-quality datasets, the optimisation of models, the establishment of an "AI preliminary recognition + expert verification" mechanism, the construction of semantic knowledge graphs, the development of linguistics-embedded models and the promotion of interdisciplinary collaboration. The advancement of the intelligent study of these manuscripts and the provision of a reference for the preservation of ancient script documents are the intended outcomes.**

**Keywords: Dunhuang Academy; Old Uighur Manuscripts; AI; Digital Protection**

## 1. Introduction

The Old Uyghur script was a significant writing system that was extensively utilised in northwestern China and Central Asia from the 9th to the 15th century. It played a crucial role in preserving a substantial corpus of manuscripts, which offer invaluable insights into the history, religion, and culture of ancient northwestern ethnic groups [1]. The region encompassing Dunhuang and its environs constituted the primary centre of Uighur activity and a pivotal locale for the retrieval of Uighur-script manuscripts. In the early 20th century, the looting of archaeological expeditions from the West resulted in the dispersion of numerous precious manuscripts overseas, with only a limited quantity being preserved in China. The Dunhuang Academy is home to the most extensive collection of Uyghur

script manuscripts in China, which can be categorised into two main types: legacy holdings and manuscripts unearthed from the northern zone of the Mogao Caves. The manuscripts under consideration encompass a wide range of texts, including Buddhist scriptures, literary and poetic works, astronomical and calendrical texts, and socio-economic records. These documents are considered to be of paramount importance for the investigation of interaction, exchange, and integration among ancient ethnic groups.

For a considerable duration, scholarly research on the Uighur-script manuscripts in the Dunhuang Academy's collection has been dominated by conventional humanities methodologies. After more than forty years of progress, transcription and collation have been accomplished for a portion of the manuscripts. However, overall advancement remains relatively restricted by severe document fragmentation, faded inscriptions, challenges in cursive decipherment, and a shortage of research personnel. Consequently, a significant proportion of manuscript fragments remain unprocessed and unpublished. Concurrently, the rapid advancement of artificial intelligence (AI) has triggered a paradigm shift in humanities and social science research. With distinct strengths in image recognition, big data analytics, and intelligent reconstruction, AI provides an innovative technical route for the conservation and investigation of ancient textual manuscripts, representing a key avenue for overcoming longstanding obstacles in Uighur script studies.

Nevertheless, the integration of AI technology with humanities research entails both opportunities and challenges. The application of AI to the identification and interpretation of understudied, extinct ancient scripts, such as the Uyghur script, as well as the interpretation of their cultural contexts and scholarly frameworks, continues to encounter substantial technical bottlenecks and difficulties. The present study commences with an examination of the provenance, developmental trajectory, and current research status of the Old Uyghur-script manuscripts preserved at the Dunhuang Academy. The present study investigates the contemporary opportunities arising from the integration of AI with Uighur script studies. In addition, it assesses the corresponding technical, academic, and practical challenges. Finally, it seeks to offer interdisciplinary perspectives for the digital preservation of ancient script manuscripts and the innovation of scholarly research. The objective is to foster the intelligent and systematic advancement of relevant studies, thereby injecting new impetus into the development of Dunhuang Studies and Old Uighur Studies.

## 2. Basic Information of Old Uighur Manuscripts Collected by Dunhuang Academy

Dunhuang Academy is one of the cultural relic conservation institutions with a relatively rich collection of Old Uighur Manuscripts in China. The collection has clear provenance and is mainly divided into two categories: Old Collection Manuscripts and Manuscripts Excavated from the Northern Area of Mogao Grottoes, each with distinctive features in origin background, preservation status and manuscript characteristics.

### 2.1 An Old Uighur Manuscripts from Dunhuang Academy Old Collection

An Old Uighur Manuscripts from Dunhuang Academy Old Collection, numbered with the prefix “D”, are precious documents collected by the Academy through various channels in its early stage. They mainly come from three sources: first, in August 1944, when the National Institute of Dunhuang Art built staff quarters, Buddhist scriptures of the Six Dynasties were found in the damaged statues of the Earth Temple built by Lama Wang in the late Qing Dynasty, laying an early foundation for the old collection; second, the free donations from collectors such as Wu Mangong, Ren Ziyi, Zhang Wei and Deng Xiufeng enriched the types and quantity of the manuscripts; third, the Academy purchased relevant manuscripts from collectors in Beijing, Lanzhou, Xi’an, Jiuquan and other places through acquisition, further improving the collection system.

This batch of documents is dominated by Chinese texts, totaling more than 1,000 items, and also includes various minority scripts such as Old Uighur, Old Tibetan, Tangut and Mongolian [2]. Among them, there are about 30 items of Old Uighur Manuscripts, which are the most numerous minority scripts in the old collection. Most of these Old Uighur Manuscripts are fragments, with some complete pages preserved. They date from the early Old Uighur remains of the Song Dynasty to the late

period (Yuan Dynasty), covering Buddhist scriptures, poems and other contents.

Professor Geng Shimin, a famous Chinese expert on Old Uighur studies, conducted the earliest academic research on An Old Uighur Manuscripts from Dunhuang Academy Old Collection. In the 1980s, he systematically studied manuscript Dy.U.D0286 for the first time, identified it as an Old Uighur translation fragment of the Avataṃsaka Sūtra in Eighty Fascicles, and published complete transliteration, Chinese translation and plate materials [3]. Relatively few relevant studies were conducted in the following three decades.

In 2015, Zhang Tieshan and Peter Zieme cooperated to study Old Uighur Manuscript Dy.U.D0707, identified it as an Old Uighur commentary on the Complete Enlightenment Sūtra, and discussed its literary relevance [4]. In 2018, the two cooperated again to decipher Dy.U.D0019 and Dy.U.D0590, and found their contents were closely related to the Pure Land School of Chinese Buddhism [5]. In 2019, they focused on five manuscripts including Dy.U.D0581V–Dy.U.D0585, completed transliteration and English translation, and elaborated on their provenance and formal characteristics [6]. In 2020, Zhang Tieshan and Peter Zieme successively published research results on Dy.U.D0198 and Dy.U.D0202, confirming they were Old Uighur fragments of the Complete Enlightenment Sūtra [7]. In the same year, Zhang Tieshan, together with Cui Yan and other scholars, conducted a special study on fragments Dy.U.D0898, Dy.U.D0848 and Dy.U.D0899, proving they were all Old Uighur translations of the Ekottara Āgama [8]. In 2021, Zhang Tieshan independently studied Dy.U.D0907, pointing out it was an Old Uighur fragment of the Comprehensive Commentary on the Abhidharmakośa-Bhāṣya [9]. In the same year, Zhang Tieshan and Peter Zieme jointly released research results on manuscripts Dy.U.D0900–Dy.U.D0910 and Dy.U.D0901–Dy.U.D0911, confirming these fragments were textually related to the Old Uighur Siddham [10]. In addition, Zhang Tieshan completed textual research on Dy.U.D0845 and Dy.U.D0904, determining they belonged to the Old Uighur Āgama system [11].

In 2022, Tusunjiang Imin and Abdurim Rustamjan published research results on Dy.U.D0746, confirming its content derived from the Old Uighur Garland of Parables of the Ten Karma Paths and speculating it might be an early Old Uighur document [12]. In 2023, the two studied Dy.U.D0886 and Dy.U.D0887, confirming they were Old Uighur fragments of the Mahāyānāparimitāyur-nāma-dhāraṇī [13]. In 2025, Abdurim Rustamjan and Tusunjiang Imin focused on manuscripts Dy.U.D0018–Dy.U.D0022, and identified poetic Old Uighur fragments of the Great White Lotus Society Sūtra through collation and textual research [14]. In the same year, Abdurim Rustamjan independently completed research on Dy.U.D1149, confirming its content came from the eighteenth chapter of the Old Uighur Diamond Sūtra [15].

In 2023, Zhang Tieshan and Aydar Mirkamali conducted textual research on Dy.U.D0200, determining it was an Old Uighur translation of the Profound Commentary on the Lotus Sūtra [16]. In 2025, Aydar Mirkamali completed research on Dy.U.D0004, finding it was related to the Old Uighur Saṃyukta Āgama [17].

In recent years, research achievements on An Old Uighur Manuscripts from Dunhuang Academy Old Collection have been increasing. More than 20 fragments have been systematically researched and verified so far. These manuscripts include various Old Uighur Buddhist scriptures and dozens of Old Uighur poetic fragments. At present, many fragments are still in the process of collation and publication, which requires more systematic arrangement and in-depth research to release the remaining manuscripts as soon as possible.

### 2.2 Old Uighur Manuscripts Excavated from the Northern Area of Mogao Grottoes

Old Uighur Manuscripts excavated from the Northern Area of Mogao Grottoes constitute the core of the Old Uighur Manuscripts collected by Dunhuang Academy. These manuscripts originated from the systematic archaeological survey of 248 caves in the Northern Area of Mogao Grottoes conducted by the Archaeological Institute of Dunhuang Academy from 1988 to 1995. The survey clarified for the first time the nature, function and distribution characteristics of the Northern Area caves, and unearthed a large number of important remains and precious cultural relics, among which documents accounted for the largest quantity.

In addition to Chinese documents, the excavated paper documents include various ethnic and

foreign scripts such as Old Uighur, Tangut, Tibetan, Mongolian, Sanskrit, Syriac and Brahmi. Among them, Old Uighur Manuscripts are the most abundant in quantity and form. Between 2000 and 2004, three volumes of archaeological reports on the Northern Area of Mogao Grottoes were published by Cultural Relics Press, yet nearly a thousand fragments (including Old Uighur fragments) remain unprocessed.

Caves in the Northern Area of Dunhuang Mogao Grottoes is a three-volume archaeological report compiled under the direction of Peng Jinzhang and Wang Jianjun, edited by Dunhuang Academy, and published by Cultural Relics Press. The report systematically collates and publishes the archaeological excavation data of the Northern Area caves, documenting in detail the location, excavation process, structure, unearthed relics, dating and functional nature of each cave in cave-number order.

The report also includes special chapters compiling multi-ethnic documents unearthed from the Northern Area, including Old Uighur, Tangut, old Tibetan and Mongolian manuscripts, together with decipherment achievements by relevant scholars. In addition to black-and-white plates, color plates are provided for some important documents to show details such as ink color and paper texture.

Volume I, published in 2000, presents archaeological findings from Caves 1 to 94 in the Northern Area. Yasen Oshur and Zhang Tieshan were jointly responsible for the collation, decipherment and translation of Old Uighur Manuscripts. In the appendix, Yasen Oshur detailed the preservation conditions of Old Uighur Manuscripts unearthed from each cave and specifically discussed Old Uighur wooden movable type [18] 352–357; Zhang Tieshan studied the Old Uighur inscriptions between lines of the Syriac document in Cave 53, providing complete Latin transliteration and Chinese translation [18].

Volume II, published in 2004, releases archaeological data from Caves 95 to 156. The collation of Old Uighur documents in this volume was completed independently by Zhang Tieshan. In the appendix, he not only described the basic information and writing features of these documents, but also provided Latin transliteration and Chinese translation for Old Uighur Manuscripts including Dy.U.B142:9, Dy.U.B154:52 and Dy.U.B156:3 [19].

Volume III was also published in 2004, comprehensively publishing archaeological data from Caves 157 to 243 and Caves 461, 463 and 464 in the Northern Area. The collation of Old Uighur documents in this volume was completed jointly by Zhang Tieshan and Yasen Oshur. In the appendix, Zhang Tieshan described the preservation status of unearthed Old Uighur Manuscripts cave by cave and conducted transliteration and translation of multiple documents [20] 383–396; Yasen Oshur systematically sorted the unearthed Old Uighur wooden movable type and successfully identified 30 specimens [20].

Since its publication, the three-volume report has provided the academic community with original materials of Old Uighur Manuscripts unearthed from the Northern Area, attracted wide attention in the field of Old Uighur studies, and promoted relevant research. In 2003, Abdurishit Yakup systematically sorted the Old Uighur Manuscripts published in Volume I, presented basic information of each document in tables, singled out Buddhist documents for thematic discussion under six categories, conducted in-depth analysis of 23 fragments, and attached transliteration and Chinese translation for selected documents [21,22].

After the full publication of Volumes II and III in 2006, he conducted a comprehensive study of all Old Uighur Manuscripts unearthed from the Northern Area. He first classified the documents into three major categories, presented core information in tables, then selected multiple documents for transliteration, translation and annotation, and conducted comparative analysis with similar manuscripts at home and abroad. Later, he continued to focus on the Northern Area Old Uighur Manuscripts, revised previous research results, and further deciphered 13 Old Uighur documents [23,24].

Many scholars at home and abroad have achieved remarkable results in the study of Northern Area Old Uighur manuscripts, including Zhang Tieshan, Aydar Mirkamali, Peter Zieme, Shōgaito Moriyasu, Dai Matsui and others.

Old Uighur Manuscripts excavated from the Northern Area of Mogao Grottoes are mostly dated to the Yuan Dynasty. Very few are completely preserved, with the majority being fragmentary. After centuries of underground burial, the paper is generally yellowed and the

writing blurred; moreover, most original copies are in cursive script, further increasing the difficulty of decipherment. More than 25 years have passed since the publication of these documents, and the academic community at home and abroad has made remarkable progress. Nevertheless, a large number of published documents still lack systematic collation, and image data of many fragments remain unopened. Therefore, the Northern Area Old Uighur Manuscripts still hold great research potential, with more breakthrough achievements expected in the future.

## 3. New Opportunities Brought by AI Technology for the Research and Preservation of Old Uighur Manuscripts Collected by Dunhuang Academy

The development of artificial intelligence has injected new vitality into the study of ancient documents and made it possible to break through the limitations of traditional methods in the preservation and decipherment of Old Uighur Manuscripts collected by Dunhuang Academy [25]. With advantages in image recognition, big data processing and intelligent analysis, AI can effectively solve problems such as scattered fragments, difficult decipherment and long collation cycles in Old Uighur studies. It provides efficient approaches in fragment assembly, character recognition, database construction and text comparison, greatly improving efficiency and accuracy, and expanding the space for tracing overseas lost documents and academic innovation, in line with the intelligent trend of ethnic linguistic and cultural research.

### 3.1 Assembly of Old Uighur Manuscript Fragments

Old Uighur Manuscripts collected by Dunhuang Academy, both from the old collection and the Northern Area excavations, show clear regional features in script, stroke, paper and ink, providing a basis for AI-based fragment assembly and provenance tracing. Traditional assembly relies heavily on personal experience, requiring manual comparison of edges, text connections and physical features, which is inefficient and prone to missing connections, especially for small, irregular fragments.

It is evident that, in accordance with the principles of deep learning, artificial intelligence (AI) has the capacity to extract the fundamental characteristics of documents that have been excavated from recognised caves in batches, thereby facilitating the establishment of a standardised feature database. For instance, through the modelling of the strokes, paper fibres and ink patterns of documents from Dy.U.B157:12 and Dy.U.B154:142, AI accurately matched them as fragments of the Old Uighur Mahāyānāparimitāyur-nāma-dhāraṇī [26]. In a similar manner, fragments of the poetic Old Uyghur Great White Lotus Society Sūtra (Dy.U.D0018–Dy.U.D0022) were successfully assembled by AI due to consistent writing style and coherent content [14]. The process that was previously known to require several months of manual labour can now be completed within a matter of days by the utilisation of artificial intelligence.

### 3.2 Recognition and Decipherment of Old Uighur Script

Old Uyghur is an extinct script which is characterised by inconsistent writing norms. The majority of manuscripts are fragmentary, cursive and eroded, thus rendering traditional decipherment highly dependent on expert experience, costly and inefficient. A study published in April 2025 demonstrated that artificial intelligence (AI) recognition of Ancient Turkic script, which is closely related to Old Uighur, has achieved a breakthrough. A convolutional neural network (CNN) model developed by Iranian and US scholars reached 96.34% accuracy on 38 characters, providing a technical paradigm for Old Uighur [27]. At present, only Dunhuang Academy in China fulfils the necessary conditions for such research. The academy houses a collection of over 1,000 Old Uyghur items, including manuscripts, stele rubbings and cave inscriptions dating from the Uyghur Khaganate to the Yuan Dynasty. This collection provides a specialist corpus. The integration of paleographers, computer scientists and humanities researchers has resulted in the formation of an interdisciplinary team.

From a technical perspective, the adoption of a hybrid framework incorporating deep learning OCR and linguistic knowledge embedding is recommended. This framework utilises CNN for the extraction of visual features, CTC for the adaptation of cursive script, and constrained language models incorporating Old Uyghur orthography and morphology to enhance accuracy.

The recommendation is to employ hyperspectral imaging and super-resolution reconstruction to standardise training data.

### 3.3 Digital Construction

The majority of Old Uyghur manuscripts are composed of paper fragments that have experienced the effects of ageing and fading over time. It is evident that traditional preservation and manual collation are inadequate for long-term conservation, due to restrictions on access and sharing that are imposed by space and time.

The integration of AI and digital technologies provides a feasible solution to this problem, by enabling the creation of a standardised and intelligent Old Uyghur database. This would facilitate digital archiving, intelligent management and global sharing. This approach is also a core technical path of Dunhuang Academy's "Digital Restoration of Overseas Lost Dunhuang Cultural Relics Project".

In digital acquisition, AI supports high-resolution scanning and intelligent image enhancement, restoring paper and ink details and completing slight character defects. This is demonstrated in the Digital "Library Cave" platform. In the context of database construction, artificial intelligence (AI) has been demonstrated to facilitate the automatic extraction of metadata, including but not limited to numerical data, excavation site details, content analysis, and research achievements. This process enables the creation of a retrievable knowledge graph, thereby enhancing the accessibility and reusability of data. The academy's Old Uyghur database has been meticulously categorised to facilitate efficient exploration through multi-dimensional queries. The D-series (old collection) and B-series (Northern Area) are now readily accessible, providing a comprehensive and interconnected framework of related studies. This initiative is poised to foster international collaboration through the facilitation of online knowledge exchange, thus promoting a global network of research and scholarship.

### 3.4 Manuscript Comparison

The research into the history of the Uyghur people frequently necessitates cross-collection and cross-version comparison in order to reveal the textual relationships and transmission routes that are of interest. The conventional approach of manual comparison is inherently inefficient and prone to overlooking subtle variations, particularly in the context of large-scale datasets.

With strong big data processing ability, AI realizes rapid comparison and accurate association mining. For example, AI detected high similarity between the Dunhuang fragment Dy.U.D1149 of the Old Uighur Diamond Sūtra and a counterpart in the Bibliothèque nationale de France, suggesting they are different editions of the same version [15]. AI reduces years of work to weeks and reveals connections overlooked by traditional methods.

In summary, AI brings all-round breakthroughs in the preservation, collation and research of Old Uighur Manuscripts collected by Dunhuang Academy, effectively solving the difficulties in fragment assembly, decipherment, collation and comparison. It promotes the intelligent, digital and international transformation of Dunhuang ancient ethnic document research and opens a new path for the inheritance of ancient ethnic languages and cultures.

## 4. Challenges of AI Technology for Old Uighur Manuscripts Collected by Dunhuang Academy

While AI technology has brought breakthroughs in the research and preservation of Old Uighur Manuscripts, its deep integration still faces numerous challenges due to the ancient nature of Old Uighur, the preservation status of the manuscripts, and the inadequate adaptation of AI in the humanities. These challenges include not only issues of precise technical adaptation but also academic difficulties in cultural interpretation and semantic explanation. Combined with cases of manuscripts unearthed from the Northern Area of Mogao Grottoes, solutions need to be explored based on the characteristics of the manuscripts and interdisciplinary advantages. The Old Uighur Manuscripts collected by Dunhuang Academy are mostly fragments of Buddhist texts, and also include letters, biographies of monks, calendars and divination texts, poems and other types. The preservation conditions and content features of these manuscripts highlight the practical difficulties in AI application.

### 4.1 High Difficulty in Old Uighur Script Recognition

As an extinct ancient script, the writing

characteristics and preservation status of Old Uighur directly increase the difficulty of AI recognition, which constitutes a core bottleneck restricting the application of the technology. Based on the cases of manuscripts collected by Dunhuang Academy, this is mainly reflected in three aspects:

First, the manuscripts are severely damaged and the recognition scenarios are complex. In particular, most Old Uighur Manuscripts unearthed from the Northern Area are fragments. After hundreds of years of underground burial and natural erosion, problems such as yellowed and broken paper, blurred writing and missing strokes are widespread. For example, the Old Uighur fragment of the Amitābha Sūtra numbered Dy.U.B53:18 retains only six lines of text on the front, with some sections seriously incomplete; blurred strokes caused by ink bleeding and incomplete page number markings on the reverse [18] Plate 36 make it difficult for AI to accurately capture complete stroke features.

Secondly, writing norms are inconsistent and script styles vary significantly. The majority of Old Uyghur copies are written in cursive script, with a high frequency of ligatures and abbreviations. Script styles vary considerably across different periods and scribes, with a paucity of unified standards. For instance, fragments of the 'Biography of Xuanzang' unearthed from the Northern Area (Dy.U.B49:41-9, Dy.U.B53:11, etc.) show both sutra script and regular script, and some fragments contain interlinear Chinese annotations. The combination of Chinese and Old Uyghur scripts poses a significant challenge to the capacity of artificial intelligence (AI) to differentiate between the two. This challenge is further compounded by the high error rate exhibited by the AI system in its recognition of cursive ligatures.

Third, training data are scarce and sample coverage is insufficient. As a minor ancient script, Old Uighur has a limited number of published clear manuscript samples. Especially for special fragments unearthed from the Northern Area (such as wooden movable-type fragments, black-paper gold-ink fragments, small-sized tiny fragments), high-quality labeled samples are even more scarce. For example, the Old Uighur wooden slip unearthed from Cave B77 is the first of its kind discovered in the Dunhuang region. The writing is blurred and there are no similar samples for AI to learn from, resulting in ineffective recognition.

In response to the above problems, combined with the characteristics of the manuscript cases, solutions can be proposed from three dimensions: First, expand high-quality training datasets. Relying on the collection advantages of Dunhuang Academy, hyperspectral imaging and super-resolution reconstruction shall be applied to published manuscripts to restore stroke details of blurred writing; meanwhile, publicly available image resources of Old Uighur Manuscripts at home and abroad shall be integrated to build a standardized recognition dataset covering samples of different scripts, preservation conditions and periods, with a focus on supplementing samples of special types of manuscripts to address the shortage of data. Second, optimize AI recognition models. Draw on the CNN model for Ancient Turkic script recognition, introduce attention mechanisms and multi-scale feature fusion techniques to enhance the model's ability to capture features such as cursive ligatures, missing strokes and intermingled scripts; through data augmentation, simulate writing forms of different damage degrees, ink bleeding and materials to improve the model's generalization ability and recognition performance on sparse texts. Third, establish a dual mechanism of "AI preliminary recognition + expert verification". Following the initial recognition stage, the data is then subjected to manual verification and correction by experts in the field of Old Uyghur. This process involves the correction of any errors that were identified during the initial recognition stage, including the correction of special terms. The corrected data is then reintegrated into the model training process, thus creating a closed loop of "recognition-verification-optimisation" that aims to balance efficiency and academic rigour.

### 4.2 Difficulties in Semantic Understanding and Cultural Context Interpretation of Old Uighur

In comparison with the technical difficulties encountered in the recognition of scripts, artificial intelligence (AI) is confronted with more significant academic challenges in terms of semantic understanding and the interpretation of cultural context in Old Uyghur manuscripts. This represents a fundamental shortcoming in the integration of AI into humanities research.

The Old Uyghur Manuscripts, as collected by Dunhuang Academy, serve as a repository for the religious beliefs, ethnic interactions and cultural heritage of the ancient Uyghur people. The majority of the contents pertain to Buddhist scriptures, commentaries and historical records, thereby embodying specific historical backgrounds and cultural connotations. Nevertheless, there is a paucity of in-depth comprehension of such humanistic information, and consequently there is an inability to accurately interpret semantics or restore contexts. This difficulty is evident in specific cases of manuscripts.

On the one hand, AI has limitations in lexical and grammatical interpretation and is prone to semantic deviation. Old Uighur no longer has living inheritance, and the semantic explanation of some words remains controversial. The texts also contain a large number of religious terms, dialectal words and loanwords, which AI can hardly interpret accurately based on existing corpora. For example, in the Old Uighur vow fragment Dy.U.B157:13, terms such as "tolulï käzdäli" (Perfect Enlightenment) and "tüzügün tuymaq yaltrïtmaklïq yol" (Wonderful Path of Universal Enlightenment) originate from the Tiantai School and Pure Land School of Chinese Buddhism. AI cannot recognize their religious connotations and can only perform literal transcription.

Another example is the colophon fragment of the Sutra on the Unthinkable Buddha Realm Spoken by Mañjuśrī numbered Dy.U.B140:5. The translator's name "Čisuin Tutung" was interpreted as "Čisang tutung" by Mr. Zhang Tieshan [28], while other scholars read it as "Čisön Tutung". If AI learns from different transcription data, it will amplify semantic misunderstanding. The "water-moon" imagery in the line "sarva suvïnta köẓüngüči ay täg burxan bolzunlar" (May all water-moons become Buddhas) in this fragment carries profound Buddhist connotations [24], but AI cannot associate this cultural context and can only translate literally. Furthermore, the capacity of artificial intelligence to differentiate between the origins and the semantic nuances of loanwords in Old Uyghur manuscripts is often deficient, which frequently results in interpretative inaccuracies.

On the other hand, AI lacks the ability to interpret cultural contexts and academic logic, making it difficult to explore the deep value of manuscripts. The creation of Old Uighur Manuscripts is closely related to specific historical scenes, ethnic interactions and religious transmission. Semantic interpretation requires interdisciplinary knowledge, which AI cannot independently associate. For instance, many fragments unearthed from the Northern Area have been revised and corrected between initial and subsequent decipherments. Such revisions require a combination of Old Uighur grammar and manuscript context, yet AI lacks such academic reasoning ability and cannot perform accurate correction.

Furthermore, some fragments of stories of Mahāsiddhas in Tibetan Buddhism mention Tibetan Buddhist figures. Their semantic interpretation requires knowledge of Tibetan Buddhist culture and the background of Uighur Buddhist transmission. AI cannot independently link the history of exchanges between the two, making it hard to uncover the value of ethnic and cultural integration behind the manuscripts. In addition, AI cannot handle textual ambiguity. For example, the Old Uighur Buddhist alliterative poems between lines of the Syriac Bible anthology numbered Dy.U.B53:14 [29,30] can only be recognized word by word by AI, which fails to understand the metrical structure and religious connotations of the poems, thus unable to meet the in-depth demands of academic research.

The core to addressing this challenge is to build an interdisciplinary integration pathway of "technology empowerment + academic leadership", realizing the deep linkage of AI with Uighur studies, Dunhuang studies, history and other disciplines. Specifically, three measures can be taken: First, construct an Old Uighur semantic knowledge graph, integrating existing lexical explanations, grammatical rules, manuscript contents and academic research findings, with a focus on incorporating special terms, loanwords and historical background information to realize the associated annotation of texts and cultural contexts, providing support for AI semantic interpretation. Second, develop a semantic interpretation model with "linguistic knowledge embedding", integrating Old Uighur orthography, morphological knowledge and cultural context information into the AI model to guide the model in inferring semantics based on context; meanwhile, introduce cross-language comparison technology to assist in interpreting ambiguous contents and religious

terms based on Chinese Buddhist classics and other Ancient Turkic manuscripts. Third, establish an interdisciplinary collaboration mechanism led by Dunhuang Academy, bringing together decipherment experts, computer technicians and literature and history scholars to jointly participate in the optimization and verification of AI models, integrating the logic and experience of humanities research into model design; meanwhile, AI shall be used to integrate massive manuscript resources and provide scholars with auxiliary tools such as semantic correlation analysis and context restoration, balancing technical efficiency and academic depth.

## 5. Conclusion

The Old Uighur Manuscripts collected by Dunhuang Academy are important historical remains of the Hexi Corridor and Central Asia from the 9th to the 15th centuries. Their preservation and research are confronted with practical problems such as scattered fragments, difficult decipherment, and long collation cycles. The application of AI technology provides an effective tool to solve these problems, yet in practical application, it is still necessary to address the particularity and complexity of Old Uighur as an ancient script.

In terms of opportunities, the transformative potential of AI technology is reflected in four aspects: first, intelligent fragment assembly, which realizes rapid and accurate matching by extracting multi-dimensional features through deep learning and has been applied in practice; second, intelligent script recognition, which can construct an OCR system suitable for Old Uighur cursive by drawing on Ancient Turkic recognition technology combined with image reconstruction; third, digital knowledge management, where an intelligent database realizes metadata extraction, multi-dimensional retrieval and cross-modal correlation of manuscripts, breaking spatial and temporal restrictions; fourth, massive manuscript comparison, which rapidly identifies version differences and textual connections, providing support for clarifying transmission routes.

I In terms of the challenges faced, two core obstacles can be identified. Firstly, the complexity of manuscript preservation and writing features, including numerous fragments, blurred writing, mixed scripts and scarce special samples, leads to insufficient AI training data. Secondly, difficulties in semantic understanding and cultural context interpretation are also identified. The rich cultural significance of Old Uyghur Buddhist manuscripts has long been recognised, yet the limitations of AI in terms of interdisciplinary knowledge association pose significant challenges to the accurate interpretation of these texts and the establishment of their historical contexts.

To address these challenges, three tasks should be prioritized in the future: first, construct a standardized Old Uighur corpus by integrating Dunhuang's collection and overseas resources and supplementing data of special types of manuscripts; second, develop a hybrid architecture of "deep learning + linguistic knowledge", embed Old Uighur knowledge into AI models, and establish a closed loop of "AI preliminary recognition – expert verification – model optimization"; third, establish an interdisciplinary collaboration mechanism to bring together scholars from various fields and integrate academic experience into model design.

The "collection–experts–institution" trinity resource allocation of Dunhuang Academy provides unique advantages for intelligent decipherment research. The deep integration of AI and Old Uighur Manuscript research can not only accelerate the collation and publication of manuscripts, but also provide a technical path for the digital restoration of Dunhuang cultural relics lost overseas, promote the intelligent and international transformation of Dunhuang Studies and Old Uighur Studies, and ultimately realize the efficient preservation and value exploration of this precious cultural heritage.

## Acknowledgements

This paper is supported by National Social Science Foundation of China Collation and Research on Old Uighur Manuscripts Collected by Dunhuang Academy (No. 21BZS025).

## References

[1] Yang, Fuxue. Uighur Documents and Uighur Culture. Beijing: Ethnic Publishing House, 2002, pp. 26–27.
[2] Tai, Huili. The Origin and Value of Dunhuang Manuscripts Collected by the Dunhuang Academy. China Calligraphy, 2019(17), pp. 50–57, 48–49.
[3] Geng, Shimin. A Study on the Fragments of the Uighur Version of the Avatamsaka Sutra

(Eighty Fascicles). Minority Languages of China, 1986(3), pp. 59–65.
[4] Zhang, Tieshan; Zieme, Peter. A Study on the Uighur Fragments of the Commentary on the Yuanjue Jing Collected by the Dunhuang Academy. Dunhuang Research, 2015(2), pp. 92–101.
[5] Zhang Tieshan, Peter Zieme, Two Old Uigur Fragments From Dunhuang Connected With The Pure Land Belief, Acta Orientalia Academiae Scientiarum Hung. Volume 71 (3), 2018, pp.253–261.
[6] Zhang Tieshan, Peter Zieme, An Old Uigur Version Of The Kasibhāradvāja Sutta Extended By A Poem, Acta Orientalia Academiae Scientiarum Hung. Volume 72 (2), 2019, pp. 179–206.
[7] Zhang Tieshan, Peter Zieme, Further fragments of the Commentary on the Yuanjue jing in Old Uigur from Dunhuang, Acta Orientalia Hung. 73 (2020) 3, pp. 439–450.
[8] Zhang, Tieshan; Cui, Yan. A Study on Three Fragments of the Uighur Ekottaragama Sutra Formerly Collected by the Dunhuang Academy. Minority Languages of China, 2020(1), pp. 3–10.
[9] Zhang, Tieshan. A Study on the Uighur Fragments of Abhidharmakośa-bhāṣya Formerly Collected by the Dunhuang Academy. Studies on the Inner Asian Languages 36, 2021.10, pp. 1–10.
[10] Zhang, Tieshan; Zieme, Peter. A Study on the Uighur Fragments of the Siddham Formerly Collected by the Dunhuang Academy. Minority Languages of China, 2021(6), pp. 106–111.
[11]Zhang, Tieshan. A Study on Two Uighur Fragments of the Āgama Sutras Formerly Collected by the Dunhuang Academy. Journal of Dunhuang Studies, 2021(1), pp. 6–11.
[12]Tusunjan, Imin; Abdurehim, Rustamjan. A Study on the Uighur Scroll of Dasakarmapathāvadānamālā Collected by the Dunhuang Academy. Journal of Dunhuang Studies, 2022(2), pp. 121–136.
[13]Tusunjan, Imin; Abdurehim, Rustamjan. A Study on the Uighur Fragments of Mahā-Amitāyur-vyākhyāna Sutra Formerly Collected by the Dunhuang Academy. Dunhuang and Turfan Studies, Vol. 22, 2023, pp. 353–365.
[14]Abdurehim, Rustamjan; Tusunjan, Imin. A Study on the Conjunction of Uighur Poetic Fragments of the Great White Lotus Society Sutra Formerly Collected by the Dunhuang Academy. Dunhuang Research, 2025(1), pp. 97–106.
[15]Abdurehim, Rustamjan. A Study on the Uighur Fragments of the Diamond Sutra Formerly Collected by the Dunhuang Academy. Collected Papers of Historical and Linguistic Studies of the Western Regions, 2025(1) (Vol. 23), pp. 163–175.
[16]Zhang, Tieshan; Aydar, Mirkamal. A Study on a Fragment of the Chinese Miaofa Lianhua Jing Xuanzan in Uighur Script Formerly Collected by the Dunhuang Academy. Dunhuang and Turfan Studies, Vol. 22, 2023, pp. 345–351.
[17]Aydar, Mirkamal. A Study on a Uighur Fragment of the Samyukta Āgama Formerly Collected by the Dunhuang Academy. Collected Papers of Historical and Linguistic Studies of the Western Regions, 2025(1) (Vol. 23), pp. 149–162.
[18]Peng, Jinzhang; Wang, Jianjun; Dunhuang Academy (eds.). Northern Caves of Mogao Grottoes at Dunhuang (Vol. 1). Beijing: Cultural Relics Press, 2000.
[19]Peng, Jinzhang; Wang, Jianjun; Dunhuang Academy (eds.). Northern Caves of Mogao Grottoes at Dunhuang (Vol. 2). Beijing: Cultural Relics Press, 2004.
[20]Peng, Jinzhang; Wang, Jianjun; Dunhuang Academy (eds.). Northern Caves of Mogao Grottoes at Dunhuang (Vol. 3). Beijing: Cultural Relics Press, 2004.
[21]Yakup,Abdurishid, On the newly unearthed Uighur Buddhist texts from thenorthern grottoes of Dunhuang.In:Sven Bretfeld and Jens Wilkens (eds.) Indienund Zentralasien:Sprach und Kulturkontakt.Vorträge des Göttinger Symposionsvom 7. Mai bis 10. Mai. (Veröffentlichungen der Societas Uralo-Altaica 61.)Wiesbaden, 2001, pp.259-276;
[22]Abdurishit Yakup. An Overview of Uighur Buddhist Documents Unearthed from the Northern Caves of Dunhuang. In: Peng Jinzhang (ed.) Studies on the Northern Caves of Mogao Grottoes at Dunhuang (Vol. 2). Lanzhou: Reader Publishing House, 2011, pp. 485–502.
[23]Abdurishid YAKUP. Uighurica from the Northern Grottoes of Dunhuang. In:

Festschrift für Shoichiro Shōnaichi: Studies on the Eurasian Languages. Tokyo: The Society for Eurasian Languages, 2006, pp. 1–41.

[24]Abdurishit Yakup. A Comprehensive Study on Uighur Documents Unearthed from the Northern Caves of Dunhuang. In: Peng Jinzhang (ed.) Studies on the Northern Caves of Mogao Grottoes at Dunhuang (Vol. 2). Lanzhou: Reader Publishing House, 2011, pp. 429–477.

[25]Cai, Mengling. Current Situation and New Opportunities: An Analysis of the Application of Artificial Intelligence in the Preservation and Restoration of Paper Documents. Journal of Archives, 2026(1), pp. 28–37.

[26]Abdurushid YAKUP, Altuigurische Aparimitāyus-Literatur und kleinere tantrische Texte, BerlinerTurfantexte XXXVI, Turnhout: Brepols Publishers n. v., 2016.

[27]Seyed Hossein Taheri at all, A Deep Learning Based Optical Character Recognition Model for Old Turkic, EAI Endorsed Transactions on AI and Robotics, Volume 4 , 2025. pp. 1-12.

[28]Zhang, Tieshan. A Study on Three Precious Uighur Buddhist Fragments Unearthed from the Northern District of Mogao Grottoes. Dunhuang Research, 2004(1), pp. 78–82.

[29]Niu, Ruji. A Restudy of the Syriac Christian-Uighur Buddhist Bilingual Manuscript Found in the Northern District of Mogao Grottoes. Dunhuang Research, 2002(2), pp. 56–63.

[30]Peter Zieme, Zwei uigurische Gedichte aus Dunhuang – Ein Deutungsversuch. Türk Dilleri Araştırmalı, Cilt 11 (2001), pp. 25-136.