# Toward Tracing Knowledge Flows in Martial Arts: Biographical Data and Interpersonal Contacts

Yumeng Hou[1,2] 

[1] Faculty of Arts and Social Sciences, National University of Singapore, Singapore
[2] Centre for Computational Social Science and Humanities, National University of Singapore, Singapore

## Abstract

This article presents an ongoing database construction effort for advancing evidence-based research on knowledge flows in Chinese martial arts. Martial arts, as embodied knowledge systems, intertwine the complexities of physical practice with ideological and sociocultural dimensions. Yet their histories remain elusive due to sparsity and divergence in documentation.

To address these challenges, we propose developing a reliable knowledge database of martial arts practitioners, with a focus on biographical information and interpersonal contacts. In doing so, we experiment with a human-in-the-loop pipeline that combines prompt engineering with domain-specific semantics, iteratively evaluated and refined by domain experts. The pipeline extracts knowledge entities from curated historical corpora, both unstructured and semi-structured, and transforms them into structured datasets.

By sharing the challenges, strategies, and preliminary outcomes, we introduce a pathway for organising knowledge within the underdocumented and heterogeneously documented martial arts historiography. This work lays a foundation for future analytics on the knowledge flows in martial arts, with potential applicability to other embodied traditions.

**Keywords:** knowledge extraction, martial arts, biographical data, interpersonal contact, prompt engineering

## 1 Introduction

Martial arts have evolved from military combat skills into civilian practices and performative traditions through a longitudinal process. The knowledge systems are rooted in embodied practices, where the body enacts, adapts, and mediates transmission through interactions with individuals, objects, natural surroundings, and social norms. For this reason, scholars have often described martial arts as multilayered microcosms of historical narrative, in which physical, technical, ideological, and sociocultural dimensions are deeply intertwined and evolve in lockstep [1; 7; 12].

Writing Chinese martial arts history, however, presents particular difficulties. On one hand, historical records of martial arts are sparse. Formal manuscripts began to appear only from the Ming dynasty onward. Yet they primarily focused on technical descriptions and were often presented in a highly compact form [14]. On the other hand, the history of martial arts in China is largely ordinary and widely distributed. Many practitioners were lower-class civilians who practised martial skills to make a living, protect their families, or safeguard their communities. They were barely literate and unable to record their own histories, nor were they considered noteworthy enough to be documented by scholars at the time.

As a result, formal documentation or biographical records of martial artists that could provide a traceable account of their practices and transmission are scarce. Although a few martial arts systems have reached a degree of consensus regarding their lineage histories, evidence is often sparse,

and different groups may preserve divergent versions of the narrative. Much of what survives has been mediated through the writings of others or passed down orally. Rather than clarifying historical truths, such sources can sometimes obscure or even complicate them.

Might history be better revealed if we integrated various sources, such as chronicles, oral histories, and anecdotal accounts, both historical and contemporary? Recent studies in computational history (e.g., [2; 5; 16]) suggest that extracting information about people's interactions and activities directly from these materials can facilitate cross-referencing and inference, potentially providing alternative forms of evidence to traditional historiography. To this end, we ask:

> *How can we pursue an evidence-based history of knowledge transmission and development in martial arts?*
>
> *How can data science and machine intelligence assist in this process?*

To address these overarching questions, we propose a ground-truth approach through the computational curation of verifiable data sourced from materials critically reviewed by a diverse cohort of martial arts scholars. By *ground-truth*, we argue for the necessity of building a robust evidentiary base from which claims about martial arts history can be tested, challenged, or refined.

In particular, when seeking to understand the transmission and development of martial arts knowledge, it is essential to examine patterns of both change and continuity, not only within individual systems but also across them. This involves investigating evidence of individual life experiences, mobility routes, kinship networks, and social interactions, and exploring how these factors influenced the knowledge transmission within specific martial arts systems.

Practitioners, especially teachers, who are typically addressed as masters in the martial arts context, are the key agents in this transmission process. While most masters pass down techniques and forms inherited through lineage, they also adapt, refine, and at times create new moves based on practical experience, stylistic preference, or pedagogical need [3; 8]. As such, developing a historically reliable database focused on these agents and the pathways of knowledge transmission, specifically their biographies and interactions, holds promise for advancing the ground-truth ideal.

This article presents a preliminary effort to develop a database representing the biographical information and interpersonal contacts of historical Chinese martial arts practitioners. Section 2 provides an analysis of the research relevance. Section 3 introduces the proposed methodology, including specific configurations and strategies, followed by a presentation of initial outcomes in Section 4. Through these sections, we discuss the challenges and explore strategies for extracting knowledge from culturally specific, underrepresented corpora in the context of Chinese martial arts. This experiment serves as a foundation for future investigations, as outlined in Section 5.

## 2   Research relevance

### 2.1   The dynamics of embodied knowledge transmission

Martial arts involve systematic clusters of principles and techniques demonstrated through choreographed forms, training methods, and fighting styles. Transmission occurs through oral instruction and physical interaction, shaped by individual experience and influenced by ritual, tradition, and the festive life of practitioner communities, forming an intrinsically sociocultural process.

In historical agrarian contexts, martial arts were typically passed down through patriarchal clan systems to ensure that core tactics and skills remained within the family to safeguard the village or community. This structure shaped teacher-disciple relationships that closely resembled kinship ties. Nonetheless, the preference for the within-clan transmission paradigm should not be conflated with the kind of secretive exclusivity dramatised in *wuxia* (literally 'martial arts and chivalry') movies and novels. In practice, many practitioners actually learned formally from multiple teachers, and informally through *qie cuo* – a practice of sparring-based exchange where techniques were

tested, confronted, and often transmitted. Through these interactions, martial artists exchange, test, adapt, and refine their skills by incorporating or responding to one another's methods [3; 11].

Additionally, while villages were generally static and martial practices often remained geographically localised, skilled practitioners were often more mobile than the average civilian. Many practitioners engaged in martial arts as part of their professional roles, with constant knowledge exchange occurring through occupational interactions and encounters. Moreover, during periods of extreme disruption such as war, natural disaster, or political suppression, the migration of villages and families also facilitated the transmission and evolution of martial knowledge along the evolving mobility networks.[1]

Such exchanges occurred not only within clans or ethnic groups but also across national boundaries. The invention of *shuangshoudao* (double-handed sword) exemplifies how local traditions studied foreign methods and developed innovations to overcome them [15]. For instance, during the Ming Dynasty *wokou* disturbances, often described as 'Japanese pirates' though in reality a multiethnic maritime force of diverse East Asian ancestry.[2] Chinese martial artists encountered foreign techniques, including the Japanese two-handed sword, which they found particularly challenging. In seeking to understand and counter these methods, they introduced innovations in both combat and weapon design, most notably the Chinese *shuangshoudao* techniques.

## 2.2  People as the contact points

As discussed, martial arts are fundamentally embodied knowledge, in which the mindful body enacts and mediates transmission through physical and intellectual interactions between individuals, between humans and objects, between people and their environment, and often, between practices and society. The human body becomes a crucial vessel for knowledge flows, where poses, gestures, and movements serve as carriers of information, representation, and expression. Practitioners, therefore, through personal interaction, act not only as agents of knowledge but as the very embodiment of *contact* in a sociocultural sense.

The concept of *culture contact* refers to the interaction between distinct cultural systems that results in exchange, adaptation, and at times conflict across technologies, languages, and practices [17]. Unlike one-directional influence, such as that imposed by colonial or feudal regimes, *contact* implies fluid and reciprocal influence. It offers a particularly valuable lens through which to understand how cultural knowledge moves across boundaries and disciplines [6]. The results of such contact may manifest in material traits, for instance, in the shape or visual characteristics of physical objects, as well as in immaterial forms, such as gestural, conceptual, or phonological features embedded within knowledge and practice.

Because martial arts, as embodied systems, evolve and spread almost exclusively through interpersonal interactions, we argue that the trajectory of *culture contact* can often be traced from the individual to the collective. When individual encounters accumulate and exert broader influence, they begin to shape what we recognise as *culture contact* at a systemic level.

Therefore, practitioners, along with the networks they form, are central to studying martial arts as a history of individual contact. These individual links, when viewed cumulatively, reveal the interactions between systems that underpin the transmission and evolution of martial knowledge. However, tracing the lives and movements of individual practitioners within Chinese martial arts is far from straightforward. Organised historical records are sparse – a condition that reflects not only

---

[1] Hakka martial arts offer a prime example. The Hakka, a Han Chinese subgroup who migrated south from the Central Plains in five major waves, developed a distinctive martial system through assimilation with local communities [4].

[2] Wang [19] discussed seven forces involved in sea raids: the Japanese of the archipelago's western rim; unlicensed merchants; seafaring bands comprised of Japanese, Chinese, Korean, European, South and Southeast Asian, and African seamen who raided and traded across East Asia; Chinese smuggler lords; residents of the western Japanese littoral from Tsushima Island to the Kii Peninsula; and sea people and water demons.

the broader historiographical challenges associated with martial texts, but also the social status of the practitioners themselves. Most were civilians rather than literati or officials, and thus rarely left behind formal biographies. What we know of them must often be pieced together from scattered sources, and their reliability depends on rigorous critical scrutiny.

In an early attempt to address the archival gap, martial historian Tang Hao compiled the *Research on Illustrated Books of Chinese Martial Arts* (《中國武藝圖籍考》) in the mid-19th century. The work remains a foundational source in modern martial arts scholarship and has since informed contemporary efforts such as the *Shedian* project (《中華射典》) and the *Collection of Rare Classical Martial Works* (《中國古代武藝珍本叢編》), among others. Moreover, more directly addressing the biographical challenge, the monumental *Dictionary of Chinese Martial Arts* (《中國武術大辭典》, hereafter '*The Dictionary*') [13], a project under the leadership of martial historian Ma Mingda, compiled concise biographical entries for over 500 notable figures throughout Chinese history, both factual and anecdotal.

*The Dictionary* stands as a landmark effort to systematise martial arts studies and serves as a foundational resource for extracting rich and reliable information in the construction of a knowledge database. It also provides a paradigm through which we aim to computationally compile and expand the resources by integrating dispersed sources that have traditionally been managed through manual scholarly processes.

## 3   Methodology

As we propose to address these research challenges through the construction of a biographical database of practitioners in Chinese martial arts history, our approach draws on principles consistent with those of the China Biographical Database (CBDB) project [9].

CBDB has proven effective in aggregating data from historical texts and reference sources to provide multiple perspectives on the lives of individuals and groups throughout Chinese history [18]. While our project takes conceptual inspiration from the CBDB paradigm, the methodology necessarily diverges. Unlike CBDB, which primarily extracts data from structured sources such as biographies and gazetteers, our work requires a more adaptable process of extraction and inference from heterogeneous and often fragmentary materials, collated and validated by a team of martial arts historians.

The traditional humanities workflow for compiling biographical profiles typically involves the following procedures:

1. Collating and validating reliable texts from multiple, distributed sources pertaining to a specific historical figure;
2. Extracting factual information through direct reference or cross-referenced inference of life events across those sources;
3. Discarding ambiguous data and selecting only verified, historically significant events to produce a succinct biographical summary, chiefly based on subjective judgment.

To enhance this process, we collaborate with martial arts scholars to emulate and also extend the traditional research pipeline through computational scalability. As described in the subsequent sections, we employ iterative prompt engineering with large language models (LLMs), in combination with an ontology framework, to extract and infer biographical information from curated and validated source texts. While algorithms handle large-scale batch extraction and inference, human experts are extensively involved in curating the corpora, adjudicating the extractions, suggesting prompt refinements – particularly for guiding the semantic interpretation of ambiguous expressions in classical Chinese – and validating the resulting datasets. Rather than generating narrative-style summaries, the output is configured as structured data entries that preserve comprehensive information while remaining operable for both human interpretation and machine processing.

## 3.1 The corpora

To construct a robust and structured dataset, we process two types of corpora that offer distinct information modalities and computational challenges.

### 3.1.1 Semistructured biographical summaries

This corpus consists of brief yet rigorously compiled biographies sourced from *The Dictionary*, written in a hybrid style that blends vernacular Chinese (*Baihuawen*) with classical Chinese (*Wenyanwen*), as shown in Figure 1(A). It covers key figures across diverse martial arts systems spanning a wide historical range, and brings a list of named entities (as in the Table of Contents). Therefore, the corpus provides a relatively structured and semantically coherent base corpus for subsequent computational processing.

### 3.1.2 Unstructured descriptive texts

The second type of corpus comprises loosely structured descriptive texts in heterogeneous linguistic styles, including classical Chinese (*Wenyanwen*) and vernacular Chinese (*Baihuawen*) (see excerpt in Figure 1(b)). These sources were collated and validated by martial arts historians as part of a preparation for composing biographies of practitioners of Yang-style Tai Chi. While all content pertaining to a single individual was grouped into a single document, their internal organisation remains unsystematised and lacks semantic markup.

**(a)** 杨兆清

　　(1883—1936)字澄甫，世以字行。河北永年人。杨式太极拳创始人杨福魁之孙。兆清身材魁梧，资质聪颖，性情和顺。幼时不甚喜拳技，年将弱冠，始从父(健侯)学拳，日夜苦练，悟拳中奥妙，技艺日精。他的拳势外软如绵，内坚如铁，动之至微，引之至长，发之至骤。他还进一步把祖传拳架修润为大架。这个拳架舒展简洁、缓慢圆活、身法中正、练法简易，促进了杨式太极拳的普及和推广。杨兆清在民国初年被聘为北京体育研究社教师，1928 年后巡回授拳于北京、南京、上海、杭州、广州、汉口等地。曾任中央国术馆武当门门长。1930 年受聘为浙江国术馆教务长。著有《太极拳使用法》、《太极拳体用全书》。
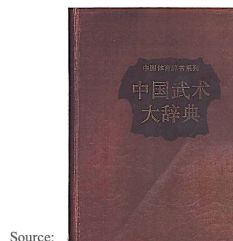
Source:

**(b)** 1.1杨式太极拳的创始和发展&杨寔[1]

　　梅兰芳演《霸王别姬》中的舞剑一场，据后来梅氏对人说，剧中一场剑舞，以前排演多次，总感不够流畅，后经杨先生一再指正，掌握不少要领，才能得心应手，运用自如。他在北京的弟子中，有名刘东汉者，原习少林拳，自以为所向无敌，轻视同侪。

　　澄甫有子四人，长振铭，次振基，三振铎，四振国，均精武术。得意门生甚多，如陈月坡、牛春明、陈微明、阎月川、王饶清、王旭东、武汇川、刘东汉、姜亭选、崔毅士、李得芳、金振华、吕殿臣、褚桂亭、邢玉臣、匡克明、奚诚甫、刘盖臣、郭荫棠、李椿年、董英杰、甫存、韩佩如、孙子玉、张钦霖、郭清杰、郑曼青、张庆麟、武志信及李亚仙等，均有专长，且各有传人。

**1.2杨澄甫[2]**

　　杨澄甫体格魁伟，生性温良敦厚，对人忠实诚恳，在武林中德高望重，从学者甚众，著名弟子有陈月坡，牛春明，陈微明，阎月川，王镜清，王旭东，刘东汉，姜亭选，武汇川,崔毅仕,李得芳,金振华,吕殿臣,褚桂亭,邢玉臣,奚诚甫,李年,董英杰,韩佩,甫存,张子玉,匡克明,张钦霖,郭清杰,郑曼青,张庆麟,武志信,傅钟文及其子杨振铭(字守中)杨振基,杨振铎，杨振国等人,盛况空前,桃李争艳,誉满全球。今天杨氏太极拳,流传世界各地,与杨澄甫的胸襟广阔,悉心授徒是分不开的。

**Figure 1:** A comparative view of two types of corpus materials related to a single practitioner entity – Yang Chengfu, a significant figure in the history of Yang-style Tai Chi. (a) A coherently written, concise biographical entry sourced from *The Dictionary*. (b) Loosely structured descriptive texts describing different facets of the practitioner's life, from various sources and registers.

## 3.2 Semantic basis

To construct a descriptive and semantically coherent database of martial arts practitioners while connecting the diverse conceptual dimensions of martial knowledge and embodied traditions, this work adopts the Martial Arts Ontology (MAon) as its semantic foundation [10].

　　MAon structures the knowledge domain of martial arts through three interrelated semantic modules: the *technical*, *stylistic*, and *social* modules. The *technical* module models the layered deployment of physical attributes in executing techniques and forms. The *stylistic* module, encompassing epistemic and symbolic dimensions, describes how technique combinations form stylistic

identities with cultural and representational significance. The *social* module addresses knowledge transmission, i.e., how martial systems are taught, learned, codified, evaluated, and disseminated.

Specifically, this work draws on the *social module,* which models different types of social agents, such as individual practitioners (class:MA_Master) and institutions (class:MA_Institute), along with their relationships to entities including places, people, objects, and time periods. Leveraging the assertions and object properties defined in this module, we extend the ontology around the MA_Master class to support a fine-grained description of individuals and their contact relations.

### 3.3 Knowledge Extraction through Prompt Engineering

Prompt engineering has gained traction as an efficient way to utilise LLMs by crafting and optimising prompts to improve task-specific reasoning. By actively engaging with the capacities of general-purpose models, while rectifying their limitations, we can guide these models to perform more effectively in disciplinary contexts. Prompt engineering thus offers a more agile and adaptable solution than fine-tuning a dedicated model, and allows for iterative refinement through the observation and refinement of the given prompt's limitations. For these reasons, at this exploratory stage, we chose to leverage prompt engineering as a rapid and iterative method for prototyping the workflow of knowledge extraction and organisation.

#### 3.3.1 Model setup

GPT-4o was selected as the base model because it demonstrated higher coherence than other models when handling undertrained, heterogeneous Chinese-language sources during our exploratory sample testing (January 2025).

#### 3.3.2 Crafting prompts

Few-shot learning is adopted as our prompting strategy. Prompts are structured as $#Identity$, $#Instruction$, followed by multiple $#Examples$, and used as a system message to generate code. This approach efficiently introduces the model to new tasks through input-output examples, allowing it to infer patterns and apply them to new data.

Two core sets of prompts have been iteratively devised (see samples in Appendix A): A.1 for extracting **practitioner entities** with explicitly identified attributes in the source texts, and A.2 for extracting both explicitly stated and implicitly inferred **contact relations** (technically, object properties) between pairs of identified practitioners. To mitigate issues of hallucination and facilitate explainable extractions, the model was instructed to append the **reference excerpt** for each relation to clarify how the inference was derived from the original text.

#### 3.3.3 Output configuration

The response for each practitioner entity extraction is configured as a JSON array comprising the following fields:

- Value properties describing the person's identity, including Name, Courtesy Name, Style Name, Aliases, Date of Birth, and Date of Death.
- Entities illustrating ethnographic information, including Ethnicity, Dynasty (or known as 'reign period'), and Place of Birth.
- Entities capturing social activities and professional engagement, including Organisations, Occupations, Works Authored, Works Mentioned, and Related Events.
  Each of these fields may contain multiple values and is therefore set as a list.
- The specific types of Martial Arts Practised by the individual.
- A brief Biography algorithmically summarised from the relevant textual sources.

Regarding contact relations, we place particular emphasis on identifying kinship, master–disciple ties, and social associations (such as colleagueship or shared place of origin or residence). The responses, also structured as JSON arrays, are designed to include informative fields that describe the relational triplet, along with the reference excerpt from the original text from which the relation was inferred. Specifically, each entry includes:

- Name and type of the start entity.
- Name and type of the end entity.
- Semantic type of the relation.
- Reference excerpt.

### 3.3.4 Normalisation

The reality of working with heterogeneous corpora is that they are rarely coherent. In the context of Chinese martial arts, this is further complicated by (1) the variation in appellations used for historical figures, (2) the complex naming conventions of martial arts systems and dynasties, and (3) the temporal ambiguity introduced by dynastic designations.

To resolve the complexity, we curated a set of standard term mapping tables by extracting entities from *The Dictionary* to regularise appellations, names of martial arts systems and those of dynasties (examples in Figure 2). During prompting, the model was instructed to reference these mappings, ensuring adherence to controlled vocabularies and avoiding confusion between distinct concepts. After extraction, relational triplets were harmonised by consolidating person entities under their original name and aligning by Gregorian calendar dates with historical dynastic periods.

**(a)**

| Name | StyleName | Alias |
|---|---|---|
| 梁学香 | | 梁学香 |
| 梁振蒲 | | 梁振蒲，估衣梁 |
| 廖观音 | | 廖观音，廖九妹 |
| 廖四公 | | 廖四公，十二峰樵 |
| 林世荣 | | 林世荣 |
| 林瑛 | | 林瑛 |
| 林尹民 | 清庵 | 林尹民，清庵，无我 |
| 刘宝珍 | | 刘宝珍，刘宝贞，飞刀刘 |
| 刘采臣 | | 刘采臣，刘差 |
| 刘德宽 | 敬远 | 刘德宽，敬远，大枪刘 |
| 刘德长 | | 刘德长，德长 |
| 刘殿琛 | 文华 | 刘殿琛，刘文华，文华，殿琛 |
| 刘短打 | | 刘短打 |
| 刘凤春 | | 刘凤春，涿州刘，翠花刘 |
| 刘奇兰 | | 刘奇兰，奇兰 |
| 刘荣庆 | | 刘荣庆，刘国庆 |
| 刘实君 | | 刘实君，刘瑞琨，快手刘 |
| 刘鬘 | | 刘鬘 |
| 刘通 | | 刘千仞，刘鬘 |
| 刘完素 | 守真 | 刘完素，守真，河间先生（常被尊称为） |
| 刘万义 | | 刘万义 |
| 刘文华 | 殿臣，殿琛 | 刘文华，文华，殿臣，殿琛 |
| 刘仙子 | | 刘仙子 |
| 刘云峰 | | 刘云峰 |
| 卢振铎 | | 卢振铎 |
| 鲁石公 | | 鲁石公 |
| 陆剑门 | | 陆剑门 |
| 陆世仪 | 道威 | 陆世仪，道威，桴亭，桴亭先生 |
| 陆游 | 务观 | 陆游，务观，放翁 |
| 吕布 | 奉先 | 吕布，奉先，飞将，温侯 |
| 吕洞宾 | | 吕洞宾，吕岩，吕祖，纯阳祖师 |
| 吕红 | | 吕红，吕短打 |
| 罗方伯 | | 罗方伯，方伯 |
| 罗光玉 | | 罗光玉 |
| 罗文源 | | 罗文源，陇上棍王之一 |
| 马超 | 孟起 | 马超，孟起 |
| 马承智 | | 马承智，马金镖 |
| 马凤图 | 健翔 | 马凤图，健翔 |
| 马贵 | 世卿 | 马贵，世卿，木马，螃蟹马 |
| 马怀德 | | 马怀德，尕爷 |
| 马金镖 | | 马金镖 |
| 马良 | 子贞 | 马良，子贞 |
| 马梅虎 | | 马梅虎 |
| 马全 | | 马全，马瑛 |
| 马全义 | | 马全义 |
| 马三元 | | 马三元 |
| 马维祺 | | 马维祺，煤马 |
| 马兴 | 鸣佩 | 马兴，鸣佩 |
| 马学礼 | | 马学礼 |
| 马英图 | 健勋 | 马英图，健勋 |
| 马玉麟 | | 马玉麟 |
| 买壮图 | | 买壮图 |
| 茅元仪 | 止生 | 茅元仪，止生，石民，逸史，东海书生，东海波臣 |

**(b)**

| Organisation |
|---|
| 武学 |
| 弓箭社 |
| 英略社 |
| 锦标社 |
| 校署 |
| 善扑营 |
| 善扑处 |
| 杆子库 |
| 义和拳 |
| 大刀会 |
| 刀客 |
| 镖局 |
| 白蜡杆会 |
| 民间拳社 |
| 拳社 |
| 武棚 |
| 中央国术馆 |
| 国际武术联合会筹备委员会 |
| 亚洲武术联合会 |
| 欧洲武术协会 |
| 中国武术协会 |
| 中国武术研究院 |
| 中国武术学会 |
| 武术辅导站 |
| 中央国立体育传习所 |
| 国立国术体育专科学校 |
| 北平体育研究社 |
| 北平国术馆 |
| 国强武术社 |
| 四民武术社 |
| 四民武术社 |
| 中国武术社 |
| 中国通省拳研究社 |
| 北平体育讲习所 |
| 河北省国术馆 |
| 天津市国术馆 |
| 中华武士会 |
| 蒲阳拳社 |
| 上海市国术馆 |
| 精武体育会 |
| 中华武术会 |
| 中华武侠会 |
| 上海第一公共体育场国术部 |
| 暑期传习所 |
| 达摩国术社 |
| 尚武进德会 |
| 致柔拳社 |
| 汇川太极拳社 |
| 武当太极拳社 |
| 上海邮务工会国术股 |

**(c)**

| Martial Arts System |
|---|
| 二十四大战拳 |
| 二十四字拳 |
| 二十四弃探马 |
| 十二短 |
| 八闪翻 |
| 三十六合锁 |
| 六步拳 |
| 内家拳六路 |
| 北拳 |
| 西家拳 |
| 西家拳 |
| 囧拳 |
| 农家拳 |
| 张飞神拳 |
| 赵家拳 |
| 童子拜观音神拳 |
| 温松钩挂拳 |
| 温家拳 |
| 霸王拳 |
| 长拳 |
| 红拳 |
| 华拳 |
| 查拳 |
| 潭腿 |
| 弹腿 |
| 唐拳 |
| 太极拳 |
| 陈式太极拳 |
| 杨式太极拳 |
| 武式太极拳 |
| 孙式太极拳 |
| 吴式太极拳 |
| 大悲拳 |
| 少林二十四炮 |
| 少林十三抓 |
| 少林八卦拳 |
| 少林五行柔术 |
| 少林五形八法拳 |
| 少林五拳 |
| 少林软开门 |
| 少林唐拳 |
| 少林禅门 |
| 心意把 |
| 地功八汉拳 |
| 地煞拳 |
| 秀拳 |
| 连拳 |
| 俞派少林拳 |
| 梅花捷拳 |
| 心意拳 |

**Figure 2:** Examples of standard term mapping tables for (a) person names, (b) organisations, and (c) martial arts systems.

## 4 Preliminary results

To assess and improve the effectiveness of the engineering pipeline, we conducted two focused case studies on biographical and relational data extraction: (1) all named practitioner entities recorded in *The Dictionary*, and (2) contemporary Yang-style Tai Chi practitioners parsed from newly collated unstructured texts, most of whom are absent from the former source. Multiple rounds of expert validation were carried out by cross-checking all extracted practitioner information, as well as kinship and master–disciple relations – the two types directly linked to knowledge transmission – for Yang-style Tai Chi, by tracing the reference excerpts back to source texts.

### 4.1 Extraction of practitioner entities

A total of 431 practitioner entities were identified from *The Dictionary*, with annotation property fields extracted where applicable (Figure 3). In addition, 43 important figures in the transmission history of Yang-style Tai Chi were retrieved from the newly curated corpus, which expands on the four individuals previously compiled in *The Dictionary*.[3]

In this exercise, prompt engineering proved readily applicable for extracting named practitioner entities from relatively structured and clean Chinese texts. The model demonstrated viable accuracy in recognising various appellations and ethnographic information of historical figures without special additional instructions in the prompts. However, it struggled to differentiate between concepts such as organisation versus dynasty, and martial arts systems versus related notions like techniques or clans. To address this issue, we compiled terminologies from *The Dictionary* and other sources into a reference file organised by category, and instructed the model to use these predefined entity groups during extraction. This refinement improved accuracy to 98%.

### 4.2 Extraction of interpersonal contacts

A total of 1,607 interpersonal contacts were identified from the integrated processing of *The Dictionary* (849) and the Yang-style Tai Chi corpus (758), of which 1,222 were validated as transmission links representing teacher-student and kinship-based knowledge-transfer relationships. Each entry was post-processed into CSV format, including the triplet information along with the reference excerpt from which the inference was drawn (Figure 4). In addition, 1,587 relations between persons and other entities (such as places or events) were identified, which could potentially expand the dataset after thorough verification.

The model was less effective in inferring *directed* interpersonal relationships than in extracting practitioner entities. While it could identify the existence of master-disciple relationships, even with relevant excerpts correctly extracted, it frequently confused their directions, i.e., reversing teacher and disciple, with initial accuracy below 50%.

Upon examining the excerpts and the original texts, we observed that the model struggled particularly with classical Chinese expressions using a single character to indicate transmission activities, such as A拜B(...师), A从B(学/练/...), A随B(学/练/...) – generally indicating that A is a student of B; or A教B, A传B, A授B – generally indicating that A is a teacher of B. To address this, prompt refinement was carried out by incorporating these inference directives into the instructions and providing additional examples for few-shot learning. Following this adjustment, the model's accuracy improved to approximately 82%, as manually verified during the expert adjudication process. While this is a promising improvement, further research is needed to enhance reliability for fully automatic extraction and to scale the approach to additional corpora.

---

[3] The 43 additional figures refer only to the focused Yang-style Tai Chi case study, one of the 644 systems compiled in *The Dictionary*. The number of entities, and thus the dataset, could increase significantly when the approach is scaled to additional styles, each requiring an expert-curated corpus and adjudication process.

# 5 Conclusion and outlook

This work explores a human-in-the-loop pipeline that leverages prompt engineering to configure more effective prompts for constructing a reliable biographical database of martial arts practitioners from scholarly curated corpora. Using this approach, we extract biographical entities with robust effectiveness and contact relations, though at a compromised performance, from heterogeneous historical materials. The JSON extractions can be converted into other formats suitable for programmatic tools, such as CSV and graph database deployment (Figure 4), to support further analytical development.

Through these experiments, we found that generally trained LLMs can be tuned via iterative prompt refinement to perform better on undertrained corpora. This approach provides a rapid prototyping solution for exploring knowledge extraction in new contexts and can significantly enhance dataset creation and curation when the data scope is manageable. However, questions of reliability and scalability remain, particularly with the presence of semantic complexity and ambiguity, before applying this approach to more comprehensive corpora for in-depth analysis.

Procedurally, expert adjudication is critical for identifying inference issues and guiding prompt refinement. Fostering this process, excerpt extraction – texts that directly underpin the model's outputs – has proven an effective practice that enhances expert efficiency while providing explainability for validated extractions.

Methodologically, we plan further research in two directions: (1) applying chain-of-thought and self-reflective instructions to probe the limits of prompt engineering for heterogeneous and context-specific Chinese corpora; and (2) developing a specialised model as a checking layer to adjudicate and correct the model's outputs, for example, by inferring and cross-checking semantic indicators of master-disciple relationships from reference excerpts.



**Figure 3:** A snapshot of extracted practitioner entities, processed from JSON into CSV format.
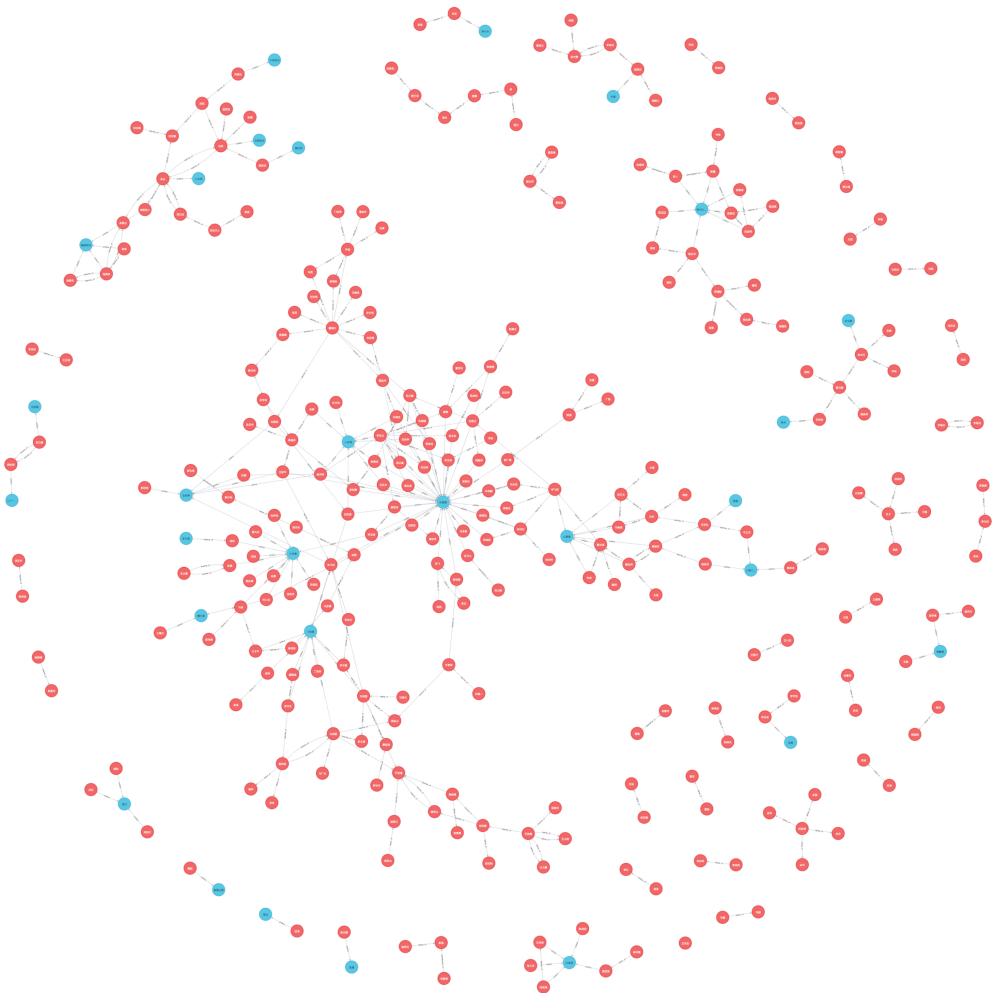
## Acknowledgements

**Figure 4:** Extraction of transmission contacts: (top) snapshot of extractions in CSV format; (bottom) preliminary deployment of the transmission network in Neo4j.

Ltd.

## References

[1]    Bowman, Paul. *Deconstructing martial arts*. Cardiff University Press, 2019. DOI: 10.18573/book1.

[2]    Cha, Javier. "To Build a Centralizing Regime: Yangban Aristocracy and Medieval Patrimonialism". In: *Seoul Journal of Korean Studies 32*, no. 1 (2019), pp. 35–80. DOI: 10.1353/seo.2019.0003.

[3]    Chao, Hing, Ma, Lianzhen, San, San, et al. *Hong Kong Martial Arts*. Ming Pao Weekly, 2014. ISBN: 9881540402, 9789881540409.

[4]    Chao, Hing, Shaw, Jeffrey, and Kenderdine, Sarah. *300 Years of Hakka Kung Fu: Digital Vision of its Legacy and Future*. International Guoshu Association, 2016.

[5]    Chen, Song and Rudolph, Henrike. "Beyond relationships and guanxi: An introduction to the research of Chinese historical networks". In: *Journal of Historical Network Research 5* (2022), pp. iii–xxxii. DOI: 10.25517/jhnr.v5i1.131.

[6]    Deagan, Kathleen, Rice, Prudence M, Schuyler, Robert L, Ramenofsky, Ann F, Schortman, Edward M, Urban, Patricia A, Hill, Jonathan D, Singleton, Theresa A, Terrell, John Edward, Stein, Gil J, et al. *Studies in culture contact: Interaction, culture change, and archaeology*. 25. SIU Press, 2015.

[7]    Farrer, Douglas S and Whalen-Bridge, John. *Martial arts as embodied knowledge: Asian traditions in a transnational world*. State University of New York Press, 2011. DOI: 10.1353/book12668.

[8]    Gotti, Roberto. "The dynamic sphere: thesis on the third state of the Vitruvian Man". In: *Martial Culture and Historical Martial Arts in Europe and Asia. Martial Studies* (2023), pp. 93–147. DOI: 10.1007/978-981-19-2037-0_4.

[9]    Harvard University, Academia Sinica, and Peking University. "China Biographical Database (CBDB)". Apr. 2019. URL: https://projects.iq.harvard.edu/cbdb.

[10]   Hou, Yumeng and Kenderdine, Sarah. "Ontology-based knowledge representation for traditional martial arts". In: *Digital Scholarship in the Humanities 39*, no. 2 (2024), pp. 575–592. DOI: 10.1093/llc/fqae005.

[11]   Lau, Kai-Yiu. "Chinese martial arts". In: *Hong Kong history: themes in global perspective*. Springer, 2021, pp. 241–260.

[12]   Lianzhen, Ma. "The Bodily Practice, Thoughts, and Beliefs in the Pre-Qin Period (before 221 BC)". In: *Routledge Handbook of Sport in China*. Routledge, 2023, pp. 9–14. DOI: 10.4324/9781003204015-3.

[13]   Ma, Xianda, Ma, Mingda, Xi, Yuntai, Gewu, Lian, Ke, Changxu, and Xuanhui, Zhang. *Dictionary of Chinese Martial Arts*. 1st ed. People's Sports Publishing House, 1990. ISBN: 7500904630.

[14]   Mingda, Ma. "A General Discussion on Establishing the Study of Ancient Martial Arts Books and Documents". In: *Discourses on the sword: Collected manuscripts*. Lanzhou University Press, 2000, pp. 315–323.

[15]   Mingda, Ma. "A Historical Study of Sword and Blade Martial Arts Exchanges Among China, Japan, and Korea". In: *Discourses on the sword: Collected manuscripts*. Lanzhou University Press, 2000, pp. 212–255.

[16] Pan, Keyao. "Networking for Historical Justice: The Application of Graph Database Management Systems to Network Analysis Projects and the Case Study of the Reparation Movement for Japanese Colonial and Wartime Atrocities". In: *Journal of Open Humanities Data 8* (2022). DOI: 10.5334/johd.76.

[17] Redfield, Robert, Linton, Ralph, and Herskovits, Melville J. "Memorandum for the study of acculturation". In: *American anthropologist 38*, no. 1 (1936), pp. 149–152. DOI: 10.1525/aa.1936.38.1.02a00330.

[18] Tsui, Lik Hang and Wang, Hongsu. "Harvesting big biographical data for Chinese history: the China Biographical Database (CBDB)". In: *Journal of Chinese History 4*, no. 2 (2020), pp. 505–511. DOI: 10.1017/jch.2020.21.

[19] Wang, Yuanfei. "Introduction: Chinese discourse of pirates and the early modern global world". In: *Writing Pirates: Vernacular Fiction and Oceans in Late Ming China*. University of Michigan Press, 2021, pp. 1–18. DOI: doi.org/10.3998/mpub.11564671.

# A Core Prompt Samples

Given the linguistic features of the corpora, the prompts consist of content written in Simplified Chinese. This section presents the prompts with English translations where necessary.

## A.1 Extracting practitioner entities

```
# Identity
You are an expert in martial studies and history. You specialise in reading classical
Chinese and modern vernacular Chinese, and in analysing historical documents.

# Instructions
Extract structured information about historical martial figures from the given text.
You must output the result as valid JSON.

## Fields to extract in JSON schema
{
  "人名": "string or null",          // Name
  "字": "string or null",            // Courtesy Name
  "号": "string or null",            // Style Name
  "别名": ["string", ...] or [],     // Aliases
  "人物简介": "string or null",       // Biography
  "民族": "string or null",          // Ethnicity
  "组织机构": ["string", ...] or [], // Organisations
  "职业": ["string", ...] or [],     // Occupations
  "朝代": "string or null",          // Dynasty
  "籍贯": "string or null",          // Place of Birth
  "生日": "YYYY-MM-DD or null",       // Date of Birth
  "卒日": "YYYY-MM-DD or null",       // Date of Death
  "创作作品": ["string", ...] or [], // Works Authored
  "提及作品": ["string", ...] or [], // Works Mentioned
  "拳种": ["string", ...] or [],     // Martial Arts Practised
  "事件": ["string", ...] or []      // Related Events
}

## Additional Rules
 //Organisations (组织机构), Occupations (职业), Works Authored (创作作品), Works Mentioned
 (提及作品), Martial Arts Practised (拳种), and
 Related Events (事件) can have multiple entries.
* Biography (人物简介) can be distilled from the original text.
* Martial Arts Practised (拳种) must use terms from the list of martial arts systems provided
 at /path/, which lists names and categorisations.
* Reign periods or vassal states are not Organisations (组织机构). Examples of organisations:
 'Yue Family Army', 'Wu Qiu Jiu', 'Seven Sages of the Bamboo Grove', and 'Tiger Ben Guard'.

# Example
<doc_input id="example-1"> /text from the Dictionary of Chinese Martial Arts/ </doc_input>
<assistant_response id="example-1">
{
  "人名": "岳飞",
  "字": "鹏举",
  "号": null,
  "别名": ["岳武穆", "岳少保"],
  "人物简介": "南宋抗金名将，精通武艺，忠勇著称。",
  "民族": "汉族",
  "组织机构": ["岳家军"],
  "职业": ["将领"],
  "朝代": "宋",
  "籍贯": "河南省安阳市",
  "生日": "1103-03-24",
  "卒日": "1142-01-27",
  "创作作品": ["武穆遗书", "满江红·怒发冲冠"],
  "提及作品": ["说岳全传"],
  "拳种": ["形意拳"],
  "事件": [ "平定剧贼陶俊、贾进和之役", "单骑攻入常州盗郭吉的军营", "刺杀金将黑风大王"]
}
</assistant_response>
```

## A.2   Extracting relations between practitioners

# Identity
You are an expert in knowledge graphs. You specialise in accurately parsing relationships
between people, events, and locations described in a document.

# Instructions
Extract structured relationship data from the given text and output the result as valid JSON.

## Relations to extract in JSON schema
{
  "起点实体": "string",     // StartEntity
  "终点实体": "string",     // EndEntity
  "起点实体类型": "string",  // StartEntityType
  "终点实体类型": "string",  // EndEntityType
  "关系类型": "string",     // RelationType
  "关系描述": "string"      // RelationDescription
}

## Additional Rules
* Focus on extracting these types of relations: between person and person, between person and
event, and between location and event.
* RelationType (关系类型) must only use defined terms from the provided list: column 'relTypes'
in 'Rel_list.csv'. If none apply, output an empty array '[]'.
* For each extracted relation, include the excerpt from the source text that explicitly
supports this relation.
* Ensure the relation forms a complete subject-verb-object statement.
*If either entity of a relation is missing, omit the relation.
* Merge mentions of the same entity across paragraphs.
* Ensure a correct relation direction. You can use indicator words to determine the direction:
  - Disciple → Master: "A拜B", "A拜师B", "A师从B", "A随B", etc.
  - Master → Disciple: "A教B", "A传B", "A指导B", "A授B", etc.

# Example
<doc_input id="example-1">
"马英图(1898—1956) 字健勋。回族。河北省沧县杨石桥(今属孟村回族自治县)人。幼从父马捷元习武"
</doc_input>

<assistant_response id="example-1">
{
  "起点实体": "马捷元",     // StartEntity: Ma Jieyuan
  "终点实体": "马英图",     // EndEntity: Ma Yingtu
  "起点实体类型": "人物",    // StartEntityType: Person
  "终点实体类型": "人物",    // EndEntityType: Person
  "关系类型": "父子",     // RelationType: Father – Son
  "关系描述":           // RelationDescription
  "马英图(1898—1956) 字健勋。回族。河北省沧县杨石桥(今属孟村回族自治县)人.幼从父马捷元习武"
}
</assistant_response>