常见中文 OCR 错误对照表

表 1：形近字混淆错误对照表

错误字符	正确字符	错误类别	形态特征	出现场景
己	己	形近字混淆	开口程度不同	"自己"误为"自已"
已	己	形近字混淆	封口程度不同	"而己"误为"而已"
己	已	形近字混淆	半开半闭	"已经"误为"己经"
曰	日	形近字混淆	扁平 vs 方正	"子曰"误为"子日"
日	曰	形近字混淆	方正 vs 扁平	"日本"误为"曰本"
入	人	形近字混淆	撇捺交叉位置	"人生"误为"入生"
人	入	形近字混淆	撇捺交叉位置	"进入"误为"进人"
术	木	形近字混淆	点的有无	"技术"误为"技木"
木	术	形近字混淆	点的有无	"树木"误为"树术"
未	末	形近字混淆	上横长短	"未来"误为"末来"
末	未	形近字混淆	上横长短	"末尾"误为"未尾"
土	士	形近字混淆	下横长短	"土地"误为"士地"
士	土	形近字混淆	下横长短	"士人"误为"土人"
天	夫	形近字混淆	顶部笔画	"天下"误为"夫下"
夫	天	形近字混淆	顶部笔画	"夫人"误为"天人"
矢	失	形近字混淆	撇的起点	"矢口"误为"失口"
失	矢	形近字混淆	撇的起点	"失去"误为"矢去"
采	彩	形近字混淆	部件差异	"采邑"误为"彩邑"
彩	采	形近字混淆	部件差异	"色彩"误为"色采"
粟	栗	形近字混淆	顶部部件	"沧海一粟"误为"沧海一栗"
栗	粟	形近字混淆	顶部部件	"栗子"误为"粟子"
己	已	形近字混淆	封口程度	"知己"误为"知已"
已	巳	形近字混淆	封口程度	"而已"误为"而巳"
乌	鸟	形近字混淆	点的有无	"乌鸦"误为"鸟鸦"
鸟	乌	形近字混淆	点的有无	"鸟雀"误为"乌雀"
免	兔	形近字混淆	点的有无	"避免"误为"避兔"
兔	免	形近字混淆	点的有无	"兔子"误为"免子"
戍	戍	形近字混淆	点的位置	"戊戍"误为"戊戍"
戍	戍	形近字混淆	点的位置	"戍守"误为"戍守"
戊	戍	形近字混淆	点的有无	"戊夜"误为"戍夜"
戍	戊	形近字混淆	点的有无	"戍边"误为"戊边"
祈	祝	形近字混淆	右部差异	"祈祷"误为"祝祷"
祝	祈	形近字混淆	右部差异	"祝福"误为"祈福"
贬	砭	形近字混淆	部首差异	"针砭"误为"针贬"
砭	贬	形近字混淆	部首差异	"贬谪"误为"砭谪"
壶	壶	形近字混淆	横线有无	"水壶"误为"水壶"
壶	壶	形近字混淆	横线有无	"壶范"误为"壶范"
赢	赢	形近字混淆	部件位置	"输赢"误为"输赢"
赢	赢	形近字混淆	部件位置	"赢利"误为"赢利"
熙	煦	形近字混淆	部件差异	"熙和"误为"煦和"
煦	熙	形近字混淆	部件差异	"煦暖"误为"熙暖"
戊	戍	形近字混淆	内部差异	"戊子"误为"戍子"
戍	戊	形近字混淆	内部差异	"戊戍"误为"戊戊"

表 2：笔画增减导致的错误对照表

错误字符	正确字符	错误类别	笔画差异	出现场景
拔	拨	笔画增减	提与竖折	"选拔"误为"选拨"
拨	拔	笔画增减	竖折与提	"拨弄"误为"拔弄"
压	庄	笔画增减	点的有无	"村庄"误为"村压"
庄	压	笔画增减	点的有无	"压力"误为"庄力"
货	贷	笔画增减	横线差异	"货物"误为"贷物"
贷	货	笔画增减	横线差异	"贷款"误为"货款"
侯	候	笔画增减	竖线有无	"王侯"误为"王候"
候	侯	笔画增减	竖线有无	"等候"误为"等侯"
治	冶	笔画增减	点的有无	"冶鍊"误为"治鍊"
冶	治	笔画增减	点的有无	"治理"误为"冶理"
竞	竟	笔画增减	横线有无	"竞然"误为"竟然"
竟	竞	笔画增减	横线有无	"竟争"误为"竟争"
茶	荼	笔画增减	横线有无	"荼毒"误为"茶毒"
荼	茶	笔画增减	横线有无	"茶水"误为"荼水"
兔	免	笔画增减	点的有无	"兔子"误为"免子"
免	兔	笔画增减	点的有无	"免除"误为"兔除"
梓	梓	笔画增减	点的位置	"桑梓"误为"桑梓"
孫	係	笔画增减	部件差异	"子孙"误为"子係"
係	孫	笔画增减	部件差异	"关系"误为"关孫"
徙	徒	笔画增减	走之差异	"迁徙"误为"迁徒"
徒	徙	笔画增减	走之差异	"徒弟"误为"徙弟"

表 3：部件混淆导致的错误对照表

错误字符	正确字符	错误类别	部件差异	出现场景
漠	漠	部件混淆	部件相似	"沙漠"误认
薄	薄	部件混淆	部件相似	"薄弱"误认
借	藉	部件混淆	简繁差异	"藉口"误为"借口"
藉	借	部件混淆	简繁差异	"借用"误为"藉用"
旗	抚	部件混淆	整体形似	历史文献中混淆
抚	旗	部件混淆	整体形似	历史文献中混淆
屹	此	部件混淆	整体形似	历史文献中混淆
此	屹	部件混淆	整体形似	历史文献中混淆
銅狄磨	銅狄摩	部件混淆	音形相似	《四库全书》整理错误
摩	磨	部件混淆	底部差异	"揣摩"误为"揣磨"
磨	摩	部件混淆	底部差异	"磨砺"误为"摩砺"
厨	厨	部件混淆	厂字头差异	"厨房"误认
厨	厨	部件混淆	厂字头差异	"厨子"误认
羡	慕	部件混淆	底部差异	"羡慕"误为"羡幕"
慕	羡	部件混淆	底部差异	"慕名"误为"羡名"
赢	羸	部件混淆	底部差异	"羸弱"误为"赢弱"
羸	赢	部件混淆	底部差异	"赢利"误为"羸利"
祭	察	部件混淆	顶部差异	"祭祀"误为"察祀"
察	祭	部件混淆	顶部差异	"观察"误为"观祭"
戍	戍	部件混淆	内部差异	"戍守"误为"戌守"
戌	戍	部件混淆	内部差异	"戊戌"误为"戊戍"
戊	戍	部件混淆	内部差异	"戊夜"误为"戍夜"

表 4：异体字与规范字对照表

异体字/原字形	规范字/正体字	错误类别	识别难度	出现场景
羣	群	异体字	中等	"群书"原写作"羣书"
群	羣	异体字	中等	"群众"原写作"羣众"
峯	峰	异体字	中等	"山峰"原写作"山峯"
峰	峯	异体字	中等	"高峰"原写作"高峯"
略	畧	异体字	中等	"简略"原写作"畧"
畧	略	异体字	中等	"策略"原写作"畧"
姊	姊	异体字	中等	"姊妹"原写作"姊妹"
姊	姊	异体字	中等	"姊姊"原写作"姊姊"
裏	里	异体字	高	"里面"原写作"裏面"
里	裏	异体字	高	"里外"可能写作"裏外"
皃	貌	异体字	高	"面貌"原写作"面皃"
貌	皃	异体字	高	"相貌"原写作"相皃"
皁	皋	异体字	高	"皋比"原写作"皁比"
皋	皁	异体字	高	"江皋"原写作"江皁"
穐	秋	异体字	高	"秋季"原写作"穐季"
秋	穐	异体字	高	"秋收"原写作"穐收"
鞠	鞠	异体字	中等	"鞠躬"多种写法
弁	弁	异体字	中等	"弁言"多种写法
彙	汇	异体字	高	"汇聚"原写作"彙聚"
汇	彙	异体字	高	"汇报"原写作"彙报"
詠	咏	异体字	中等	"歌咏"原写作"歌詠"
咏	詠	异体字	中等	"咏叹"原写作"詠嘆"
羣	群	异体字	中等	"群众"原写作"羣众"
粧	妆	异体字	高	"妆奁"原写作"粧奁"
妆	粧	异体字	高	"梳妆"原写作"梳粧"
牀	床	异体字	中等	"床榻"原写作"牀榻"
床	牀	异体字	中等	"床位"原写作"牀位"
竝	並	异体字	高	"并且"原写作"竝且"
並	竝	异体字	高	"合并"原写作"合竝"
話	话	异体字	低	"说话"繁体写作"說話"
說	说	异体字	低	"说明"繁体写作"說明"
國	国	异体字	低	"国家"繁体写作"國家"
國	国	异体字	中等	民间俗写"国"
荳	豆	异体字	高	"豆蔻"原写作"荳蔻"
豆	荳	异体字	高	"豆腐"原写作"荳腐"

表 5：通假字识别错误对照表

OCR 识别结果	可能本字	错误类别	通假说明	出现场景
早	蚤	通假字	音近通假	"蚤"通"早"，"旦暮"可能写作"蚤暮"
蚤	早	通假字	音近通假	"蚤"通"早"，但 OCR 可能误识
反	返	通假字	音同通假	"返"通"反"，"反国"即"返国"
返	反	通假字	音同通假	"反"通"返"，OCR 需语境判断
知	智	通假字	音同通假	"智"通"知"，"知者"即"智者"
智	知	通假字	音同通假	"知"通"智"，古籍常见
内	纳	通假字	音近通假	"纳"通"内"，"内之"即"纳之"
纳	内	通假字	音近通假	"内"通"纳"，语境判断困难
辟	避	通假字	音同通假	"避"通"辟"，"辟人"即"避人"
避	辟	通假字	音同通假	"辟"通"避"，OCR 需语境判断
女	汝	通假字	音同通假	"汝"通"女"，"女知之"即"汝知之"
汝	女	通假字	音同通假	"女"通"汝"，第二人称代词
见	现	通假字	音同通假	"现"通"见"，"见在"即"现在"
现	见	通假字	音同通假	"见"通"现"，表示显露
莫	暮	通假字	音同通假	"暮"通"莫"，"莫春"即"暮春"
暮	莫	通假字	音同通假	"莫"通"暮"，时间词
属	嘱	通假字	音同通假	"嘱"通"属"，"属予"即"嘱予"
嘱	属	通假字	音同通假	"属"通"嘱"，嘱托之意
厌	餍	通假字	音同通假	"餍"通"厌"，"厌足"即"餍足"
餍	厌	通假字	音同通假	"厌"通"餍"，满足之意

表 6：标点符号识别错误对照表

OCR 识别结果	正确标点/内容	错误类别	错误原因	出现场景
。	无	标点错误	古籍无标点，OCR 误加	古籍原文无句读
，	无	标点错误	古籍无标点，OCR 误加	古籍原文无句读
、	无	标点错误	古籍无标点，OCR 误加	古籍原文无句读
。	。	标点错误	位置错误或遗漏	断句位置不当
。	，	标点错误	标点类型混淆	句号与逗号混淆
，	。	标点错误	标点类型混淆	逗号与句号混淆
？	。	标点错误	反问句标点错误	反问句末误用句号
！	。	标点错误	感叹句标点错误	感叹句末误用句号
：	、	标点错误	冒号与顿号混淆	领起词后误用
；	，	标点错误	分号与逗号混淆	分句之间误用
「	"	标点错误	引号类型差异	古籍引号与现代不同
」	"	标点错误	引号类型差异	古籍引号与现代不同
『	"	标点错误	引号类型差异	古籍引号与现代不同
』	"	标点错误	引号类型差异	古籍引号与现代不同
阙文	○	符号错误	古籍符号误认	古籍用○表示阙文
阙文	□	符号错误	古籍符号误认	古籍用□表示阙文
·	、	符号错误	古今符号差异	古籍句读符号
、	·	符号错误	古今符号差异	古籍句读符号

表 7：印章和页面元素导致的识别错误对照表

OCR 识别结果	正确内容	错误类别	错误原因	出现场景
印章文字	正文文字	页面元素混淆	印章叠印在正文字上	印章覆盖正文区域
眉批文字	正文文字	页面元素混淆	眉批位置判断错误	页眉区域
正文文字	眉批文字	页面元素混淆	眉批误认为正文	页眉区域
夹注文字	正文文字	页面元素混淆	夹注误入正文	行间小字
正文文字	夹注文字	页面元素混淆	正文误认为夹注	行间小字附近
页码文字	正文文字	页面元素混淆	页码误入正文	页面边缘
书眉文字	正文文字	页面元素混淆	书眉误入正文	页面上方
版心文字	正文文字	页面元素混淆	版心误入正文	页面中间折缝
墨渍	字符	污渍误认	墨渍形似字符	页面污渍区域
字符	墨渍	字符误删	字符误认为污渍	字迹模糊区域
水渍	字符	污渍误认	水渍形似字符	页面水渍区域
虫蛀	字符	缺损误认	虫蛀形似字符	虫蛀区域
字符缺失	虫蛀	字符误删	字符区域误判为虫蛀	虫蛀区域附近

表 8：竖排格式导致的识别错误对照表

错误类型	具体表现	错误原因	示例
行序颠倒	上下行顺序错误	竖排行序识别错误	先读下行后读上行
字序错误	单字上下颠倒	单字方向判断错误	"子"读为"孑"
行混入	他行文字混入	行间距离过近	A 行末字与 B 行首字混淆
跨行字符	单字分跨两行	字符分割错误	大字被切分到相邻行
注释混入	夹注误入正文	夹注位置判断错误	小字夹注被识别为大字正文
正文遗漏	夹注区域正文被遗漏	正文被误判为夹注	夹注附近的大字被误判
标题误认	标题字误为正文	标题格式判断错误	章节标题被识别为正文第一句
正文误认	正文误为标题	标题格式判断错误	正文首句被误判为标题

表 9：相似偏旁部首混淆错误对照表

错误字符	正确字符	涉及部首	混淆原因	出现场景
漠	漠	氵 vs 氵	同一部首不同写法	"沙漠"的两种写法
清	清	氵 vs 氵	点画位置差异	"清水"的不同写法
决	决	冫 vs 氵	两点水与三点水	"决定"vs"决定"
决	决	氵 vs 冫	三点水与两点水	"决定"vs"决定"
冷	冷	冫 vs 冫	两点水不同写法	"寒冷"的不同写法
凉	凉	冫 vs 氵	两点水与三点水	"凉爽"vs"凉爽"
凉	凉	氵 vs 冫	三点水与两点水	"凉爽"vs"凉爽"
况	况	冫 vs 氵	两点水与三点水	"况且"vs"况且"
况	况	氵 vs 冫	三点水与两点水	"况且"vs"况且"
草	艸	艹 vs 艸	草字头不同写法	"草木"的不同写法
花	華	艹 vs 艹	草字头简化程度	"花"与"華"
華	花	艹 vs 艹	草字头简化程度	"華"与"花"
群	羣	君 vs 君	部件位置差异	"群众"vs"羣众"
羣	群	君 vs 君	部件位置差异	"羣众"vs"群众"
峰	峯	山 vs 山	山字旁位置差异	"高峰"vs"高峯"
峯	峰	山 vs 山	山字旁位置差异	"高峯"vs"高峰"
略	畧	田 vs 田	田字旁位置差异	"简略"vs"畧"
畧	略	田 vs 田	田字旁位置差异	"畧"vs"简略"
鹅	鵝	我 vs 我	左右结构差异	"天鹅"vs"天鵝"
鵝	鹅	我 vs 我	左右结构差异	"天鵝"vs"天鹅"
拿	拏	手 vs 手	手字旁位置差异	"拿捏"vs"拏捏"
拏	拿	手 vs 手	手字旁位置差异	"拏捏"vs"拿捏"

表 10：古籍特殊字符识别错误对照表

OCR 识别结果	正确内容	错误类别	错误原因	出现场景
□	缺字	无法识别	字符破损或罕见	古籍残损处
	缺字	编码错误	字符不在 Unicode 中	罕见异体字
乱码	生僻字	编码错误	字符编码问题	罕见古籍用字
空	空白	避讳字	避讳缺笔或不写	皇帝名讳
玄	玄(缺末笔)	避讳字	避讳缺笔	清代避康熙讳
玄(缺末笔)	玄	避讳字	缺笔还原错误	清代文献避讳还原
真	真(缺末笔)	避讳字	避讳缺笔	清代避雍正讳
真(缺末笔)	真	避讳字	缺笔还原错误	清代文献避讳还原
民和	民和(缺笔)	避讳字	避讳缺笔	清代避乾隆讳
寧	寧	避讳字	避讳改字	清代避讳改字
曆	歷	避讳字	避讳改字	清代避乾隆讳
歷	曆	避讳字	避讳字还原错误	清代文献处理

表 11：古籍常见讹误类型及 OCR 识别对照表

讹误类型	原文/正确	讹误/OCR 误认	错误原因	文献依据
形近致误	愦	慨	形近字混淆	"慨"与"愦"形近而讹
形近致误	概	慨	形近字混淆	"慨"与"概"形近而讹
音同致误	以	已	音同字混用	古籍常见音同替代
音近致误	铜	銅	繁简差异	OCR 繁简转换错误
形近致误	砭	贬	形近字混淆	"针砭"误为"针贬"
形近致误	粟	栗	形近字混淆	"沧海一粟"误为"沧海一栗"
音形相似致误	摩	磨	音形相似	"铜狄摩"误作"铜狄磨"
笔画增减致误	荼	茶	笔画差异	"荼"多一横
部件混淆致误	戍	戌	内部部件差异	"戍"与"戌"形近
异体字致误	羣	群	异体字误认	OCR 未能识别异体
通假字致误	蚤	早	通假字误判	"蚤"通"早"
避讳字致误	玄(缺笔)	玄	避讳缺笔	清代避康熙讳

以下为更多常见 OCR 错误对照示例，供实际应用参考：

附录表 1：高频形近字错误对照

错误	正确	错误	正确	错误	正确	错误	正确
已	己	巳	己	曰	日	入	人
术	木	末	未	士	土	夫	天
矢	失	彩	采	栗	粟	鸟	乌
兔	免	戍	戌	祝	祈	贬	砭
壶	壶	羸	赢	煦	熙	拨	拔
庄	压	贷	货	候	侯	冶	治
竟	竞	荼	茶	徒	徙	藉	借

附录表 2：高频异体字对照

异体	正体	异体	正体	异体	正体	异体	正体
羣	群	峯	峰	署	略	姊	姊
裏	里	皃	貌	皋	皋	龝	秋
彙	汇	詠	咏	粧	妆	牀	床
竝	並	荳	豆	傢	家	伕	夫
佈	布	佔	占	伬	尺	仿	仿
夠	够	姏	妊	姍	姗	姪	侄

附录表 3：高频避讳字对照

避讳字	原字	避讳对象	避讳方式	备注
玄(缺末笔)	玄	康熙帝	缺笔	缺最后一笔
曆	歷	乾隆帝	改字	"歷"改为"曆"
寧	寧	道光帝	改字	"寧"缺笔或改写
胤	胤	雍正帝	缺笔或改字	"胤"字避讳
禛	禛	雍正帝	缺笔或改字	"禛"字避讳
弘	弘	乾隆帝	缺笔或改字	"弘"字避讳