生僻字与字符编码
生僻字为什么打不出、显示不了： Unicode 汉字区、 GB 18030 、乱码成因、输入方法、异体字和变体选择符、私用区，以及发现缺字怎么反馈。

生僻字为什么打不出、显示不了
Unicode 里的汉字：基本区和扩展区
GB 18030：中文编码的国家标准
乱码从哪来： GBK、 Big5 和 UTF-8
生僻字怎么输入
异体字、变体选择符和汉字描述序列
网页和系统里怎么显示生僻字
私用区（ PUA）里的字为什么靠不住
人名、地名里的生僻字：系统怎么做
发现缺字怎么反馈

生僻字为什么打不出、显示不了

> 适合谁看：名字、地名里有生僻字，在系统里打不出、显示成方框或问号，想弄清原因的人。

一个字要「走通」三关

一个字从输入到显示，要过三关：

1. 有编码：这个字在 Unicode / GB 18030 里有编号吗？
2. 能输入、能存：输入法能打出来吗？软件、数据库能把它完整存下来吗？
3. 有字形：当前用的字体里有这个字的字形吗？

任何一关不通，都会出问题。

看症状判断卡在哪一关

症状	最可能的原因	卡在哪
显示成空心方框「□」或带编号的小方块	字体里没有这个字	第 3 关
显示成「?」或「�」	存储或传输时编码出错，字已经丢了	第 2 关
显示成乱七八糟的其他字（乱码）	编码对不上，例如 GBK 当成 UTF-8 读	第 2 关
输入法打不出	输入法词库里没有，或这个字还没有编码	第 1 或第 2 关
同一段文字里这个字的字体明显不一样	主字体缺字，系统自动换了备用字体	第 3 关（已经补救）

方框和问号的区别很重要：

方框说明字还在，只是没字形，换个字体就能显示；问号说明字已经丢了，换字体也救不回来。

为什么生僻字特别容易出问题

- 编码靠后：常用汉字都在 Unicode 的基本区；很多生僻字在后来增加的扩展区，位于「辅助平面」，有些老软件、老数据库处理不了，

见《 Unicode 里的汉字：基本区和扩展区》 ；
字体不全 ：多数字体只覆盖常用的几千到两万多字，扩展区的字很少有字体收录，见
《一款中文字体要做多少字》 ；
还没编码 ：极少数字还没有进入标准，只能临时用私用区或图片，见 《私用区（ PUA
）里的字为什么靠不住》 。

下一步看哪里

想自己打出来：见 《生僻字怎么输入》 ；
想让网页或系统正常显示：见 《网页和系统里怎么显示生僻字》 ；
做业务系统、要存人名：见 《人名、地名里的生僻字：系统怎么做》 ；
显示成方框的通用排查：见 《繁体、日文、生僻字显示成方框怎么办》 。

Unicode

里的汉字：基本区和扩展区

> 适合谁看：想知道一个字的「编码」是什么、为什么有的字在「扩展区」、扩展区的字为什么难处理的人。

B

码位：每个字的编号

Unicode 给每个字符一个编号，叫码位，写作 `U+` 加十六进制数，例如「中」是 U+4E2D。

Unicode 把码位分成 17 个「平面」，每个平面 65,536 个位置：

- 第 **0** 平面（基本多文种平面）：最常用的字符，包括常用汉字；
- 第 **2**、**3** 平面：几乎全是汉字扩展区；
- 第 15、16 平面：私用区，见《私用区（PUA）里的字为什么靠不住》。

汉字分布在哪些区块

下表按 Unicode 17.0（2025 年）的数据统计：

区块	码位范围	汉字数	所在平面
中日韩统一表意文字（基本区）	4E00–9FFF	20,992	0
扩展 A	3400–4DBF	6,592	0
扩展 B	20000–2A6DF	42,720	2
扩展 C	2A700–2B73F	4,160	2
扩展 D	2B740–2B81F	222	2
扩展 E	2B820–2CEAF	5,774	2
扩展 F	2CEB0–2EBEF	7,473	2
扩展 I	2EBF0–2EE5F	622	2

区块	码位范围	汉字数	所在平面
扩展 G	30000–3134F	4,939	3
扩展 H	31350–323AF	4,192	3
扩展 J	323B0–3347F	4,298	3
兼容区中的 12 个统一汉字	F900–FAFF	12	0

合计 101,996 个统一汉字。 2026 年发布的 Unicode 18.0 又在扩展 D 补了 1 个。

扩展区为什么难处理

- 在辅助平面：第 0 平面以外的字，在 UTF-8 里占 4 个字节，在 UTF-16（Windows、Java、JavaScript 内部常用）里要用两个单元（「代理对」）表示。只按 3 字节或 1 个单元处理的老程序会出错；
- 字体少：扩展 B 以后的字，多数常用字体都没有；
- 输入法少：很多输入法默认不出扩展区的字。

「统一」是什么意思

中国大陆、台湾、香港、日本、韩国、越南的汉字，如果只是写法细节不同，会合并成一个码位，由字体决定显示哪种写法，见《同一个字，为什么各地写法不同》。真正不同的字（如简繁）才分开编码。

怎么查一个字的编码

- Unicode 官方汉字数据库（Unihan）：输入字或码位查询；
- 在电脑上：Windows 的「字符映射表」、macOS 的「字符检视器」都能显示码位。

GB 18030

：中文编码的国家标准

> 适合谁看：做软件、做政务和公共服务系统，需要了解 GB 18030 是什么、要满足哪个级别的人。

它是什么

GB 18030《信息技术 中文编码字符集》是中国的强制性国家标准，规定了汉字和多种少数民族文字的编码。它和 Unicode 一一对应，可以理解为「中国版的 Unicode 编码方式」，同时兼容老的 GB 2312 和 GBK。

- 2000 年首次发布，2005 年第一次修订；
- 现行版本是 **GB 18030-2022**，2022 年 7 月发布，2023 年 8 月 1 日实施；
- 共收录汉字 87,887 个，比上一版增收 1.7 万余个生僻汉字，收录了《通用规范汉字表》全部汉字，覆盖绝大部分人名、地名用生僻字。

三个实现级别

标准第 9 章规定了三个实现级别：

级别	范围	谁要满足
级别 1	单字节、双字节编码部分，以及 CJK 统一汉字（基本区）和扩充 A	所有适用的产品
级别 2	级别 1，加上《通用规范汉字表》中不在级别 1 里的编码汉字	系统软件和支撑软件（操作系统、数据库、中间件等）
级别 3	级别 2，加上本标准规定的全部汉字和康熙部首	用于政务服务和公共服务的产品

标准注明的政务服务和公共服务行业包括铁路、道路、水上、航空运输，邮政，金融，保险，卫生，国家机构，社会保障等。

第 1 号修改单
2024 年 9 月 29 日发布的第 1 号修改单，把 Unicode 后来新增的扩展 G、 H、 I 区等汉字加入级别 3 ，合计新增 9793 个汉字；级别 1 和级别 2 的范围没有变化。

对字体和软件意味着什么
级别 3 是「移动目标」 ： Unicode 继续增加扩展区，标准也可能继续跟进；
单款字体很难包揽 ：几万个汉字通常要靠多款字体配合，系统里用「备用字体」补字，见
《网页和系统里怎么显示生僻字》 ；
软件要能处理 4 字节字符 ：见 《人名、地名里的生僻字：系统怎么做》 ；
私用区不再需要 ：老版本要求的部分私用区映射，在 2022 版中已经改为标准码位，见 《私用区（ PUA）里的字为什么靠不住》 。

是否需要做产品认证、具体合规要求，请以标准原文和认证机构的规定为准。

乱码从哪来：GBK、Big5 和 UTF-8

> 适合谁看：打开文件、网页、表格时遇到乱码，想知道原因和解决办法的人。

字符集和编码方式

- 字符集：收哪些字、每个字什么编号（如 Unicode、GB 18030）；
- 编码方式：把编号变成字节存起来的规则（如 UTF-8、UTF-16、GBK）。

同一个「中」字，用 UTF-8 存是 3 个字节，用 GBK 存是 2 个字节。读的时候用错规则，就会得到乱码。

常见的中文编码

编码	范围	说明
GB 2312	6763 个汉字	早期简体中文编码
GBK	两万多个汉字	GB 2312 的扩展，Windows 简体中文系统长期默认使用
GB 18030	全部 Unicode 字符	国家标准，兼容 GBK，见《GB 18030：中文编码的国家标准》
Big5	繁体常用字	台湾、香港早期常用
UTF-8	全部 Unicode 字符	网页和现代软件的事实标准

典型乱码现象

- 「锟斤拷」：一段文字曾经被替换成了替代字符「�」，再用 GBK 显示出来；原文已经丢失，无法还原；
- 「浣犲ソ」这类怪字：UTF-8 的文字被当成 GBK 读了，用正确的编码重新打开就能恢复；

CSV 用 Excel 打开乱码 ：文件是 UTF-8 编码， Excel 按系统默认编码打开。可以在 Excel
中用「数据 → 从文本 /CSV」导入并选择 UTF-8，或保存为带 BOM 的 UTF-8。

怎么避免
1. 统一用 UTF-8：网页、数据库、接口、文件都用 UTF-8；
2. 网页声明编码 ： <meta charset="utf-8"> 写在 <head> 最前面；

3. 数据库用完整的 UTF-8：例如 MySQL 要用 utf8mb4，见 《人名、地名里的生僻字：系统怎么做》
；
4. 和老系统交换数据时 ：明确约定编码，需要中文国标编码时用 GB 18030 而不是 GBK，GBK
存不下扩展区的生僻字。

乱码和方框不是一回事

乱码是 编码 问题，换字体没用；方框是 字体缺字 ，换字体就好。区别见 《生僻字为什么打不出、显示不了》 。

生僻字怎么输入

适合谁看：不会读、打不出某个生僻字，或者需要在电脑和手机上输入人名、地名用字的 “
人。

方法一：输入法的笔画和拆字输入
不知道读音时，很多拼音输入法提供「笔画」和「拆字」输入。以搜狗输入法的「 U
模式」为例（以官方帮助为准）：

笔画输入 ：先按 u，再按笔顺输入笔画： h 横、 s 竖、 p 撇、 n 捺（点）、 z
折。例如「木」输入 uhspn；
拆字输入 ：先按 u，再依次输入各部分的拼音。例如「林」拆成两个「木」，输入 umumu
；也可以按部首拆，如「氻」拆成「氵」和「力」；
双拼用户用 Shift + U 进入这个模式。

其他输入法通常也有类似功能，名称和按键可能不同，查看各自的帮助。手机输入法有的需要先开启「生僻字
」相关选项。

方法二：系统自带的字符工具
Windows「字符映射表」 ：开始菜单搜索「字符映射表」，选择字体后可按 Unicode
码位查找、复制；
macOS「字符检视器」 ：在任意输入框按 Control + Command + 空格
，可以搜索、浏览汉字，查看码位。

前提是你选的字体里有这个字。

方法三：知道码位，直接输入

如果查到了这个字的 Unicode 码位（查法见 《Unicode 里的汉字：基本区和扩展区》 ）：

Microsoft Word：输入码位（如 20B9F），紧接着按 Alt + X，就会变成对应的字；

- **macOS**：在「系统设置 → 键盘 → 输入法」里添加「Unicode 十六进制输入」，按住 `Option` 键输入码位；
- 网页 / 程序里：HTML 写 `𠮟`，CSS 写 `\20B9F`，JavaScript 写 `"\u{20B9F}"`。

方法四：复制粘贴

从权威字典网站或官方数据库复制，再粘贴。注意：

- 复制前确认是正式编码的字，而不是图片或私用区字符，见《私用区（PUA）里的字为什么靠不住》；
- 粘贴后显示方框，说明字已经进来了，只是当前字体没有，见《网页和系统里怎么显示生僻字》。

实在没有编码怎么办

极少数字还没有进入 Unicode。临时办法：

- 在正式文书中，按相关部门的要求处理；
- 一般场合，可以用「𫝀 力」这样的描述序列说明字形，见《异体字、变体选择符和汉字描述序列》。

异体字、变体选择符和汉字描述

序列

适合谁看：遇到「同一个字但写法不一样，又必须按某种写法显示」的需求，比如户籍、 “
古籍、日本人名的人。

异体字
读音和意思相同、写法不同的字叫异体字，例如「峰」和「峯」。

如果写法差别足够大， Unicode 会给它们 不同的码位 ，可以直接输入；
如果差别很细微， Unicode 认为是同一个字，只给 一个码位 ，具体写法交给字体。见

《同一个字，为什么各地写法不同》 。

第二种情况下，如果业务 必须
显示特定写法（例如人名证件要和登记的字形一致），就要用到「变体选择符」。

汉字变体序列（ IVS）
做法是在字后面紧跟一个看不见的「变体选择符」（码位 U+E0100 到 U+E01EF
），告诉字体「请显示这个字的第几种写法」：

基础字 + 变体选择符 = 一个指定的字形 “

这些组合需要在 Unicode 的「汉字变体数据库」（ IVD）中登记；
已登记的集合包括 Adobe-Japan1、Hanyo-Denshi、Moji_Joho、MSARG、KRName、CAAPH
等，其中 Moji_Joho 集合最近一次在 2026 年 8 月增补；
多数集合服务于日本的人名和行政用字，也有服务于其他地区的集合；在中国大陆的日常中文场景中
用得还不多。

使用 IVS 的前提
1. 字体支持 ：字体里要有对应的变体字形和映射；不支持的字体会忽略选择符，显示默认字形；
2. 软件支持 ：排版软件、浏览器要能正确处理（主流浏览器和操作系统一般都支持）；
3. 数据要保留
：复制、存储、截断字符串时不能把选择符丢掉；按「字符数」限制长度时，选择符也会占位置。

汉字描述序列（ IDS）
有些字没有编码，或者想向别人说明一个字的结构，可以用「汉字描述字符」描述，例如：

⿰氵力 表示左「氵」右「力」，即「氻」；
⿱艹田 表示上「艹」下「田」。

常用的描述字符： ⿰ 左右、 ⿱ 上下、 ⿲ 左中右、 ⿳ 上中下、 ⿴ 全包围、 ⿵ 上三包围等。 Unicode 15.1 又增加了几个描述字符。

注意： 描述序列只是说明，不会自动变成那个字

，显示出来就是几个符号排在一起。适合用来记录、沟通、提交缺字反馈，见 《发现缺字怎么反馈》 。

网页和系统里怎么显示生僻字

适合谁看：网站、 App “
或文档里有生僻字显示成方框，想让它正常显示的开发者和设计师。

先确认是「缺字形」而不是「丢了字」
显示成方框，说明字还在，只是字体里没有；显示成问号，说明存储或传输时已经丢了，要先修编码。判断方法见 《生僻字为什么打不出、显示不了》 。

系统怎么补字
操作系统和浏览器都有「字体回退」机制：主字体里没有某个字时，会自动去其他已安装的字体里找。 Windows 和 macOS
都自带一些覆盖扩展区的字体，所以同一个生僻字，在电脑上常常能显示，只是字体风格和周围不一样；在字体较少的设备（部分手机、嵌入式设备）上就可能是方框。

网页：在字体栈后面加一款「补字字体」

body {
font-family:
"主字体", /* 你的正文字体 */
"Plangothic P1", /* 补字字体：覆盖扩展区 */ "Plangothic P2",
sans-serif;
}

要点：

补字字体放在主字体后面 ：主字体有的字用主字体，没有的才用补字字体；

- 字体名以实际文件为准：上例中的遍黑体分成 P1、P2 两个文件，名称请以下载的文件和项目说明为准；
- 用 `unicode-range` 按需下载：把补字字体的 `@font-face` 限定在扩展区码位，页面里没有这些字时浏览器就不会下载：

```
@font-face {
  font-family: "补字字体";
  src: url("/fonts/fallback-extb.woff2") format("woff2");
  unicode-range: U+20000-2A6DF, U+2A700-2EE5F, U+30000-3347F;
}
```

- 风格尽量接近：正文是黑体就配黑体风格的补字字体，避免一个字突兀；
- 注意许可：补字字体也要允许网页嵌入，见《自托管网页字体的许可要点》。

字体栈的一般写法见《中文字体栈怎么写》。

文档和设计稿

- Word、WPS 等会自动回退到系统字体；需要统一风格时，手动给生僻字指定一款有这个字的字体；
- 导出 PDF 时嵌入字体，接收方才能看到同样的字；
- 设计稿交给前端时，把用到的生僻字列出来，确认网页字体里有没有，见《从设计稿到上线：怎么和前端对接字体》。

用文风网页字体时

文风提供的网页字体以各款字体本身收录的字为准，没有收录的字会由你在字体栈里写的后备字体显示。所以生僻字多的页面，同样建议在字体栈后面加补字字体。网页字体的接法见《用文风网页字体：一页看懂正确接法》。

私用区（ PUA

）里的字为什么靠不住

适合谁看：手里有「自定义字」「补字」文件或老系统数据，发现换个电脑、换个字体字 “
就变了或没了的人。

什么是私用区
Unicode 留出了一些码位，不分配给任何字符，供个人或机构自己约定使用，叫「私用区」：

范围 位置数

U+E000–U+F8FF 6,400（基本平面）

U+F0000–U+FFFFD 65,534（第 15 平面）

U+100000–U+10FFFD 65,534（第 16 平面）

过去，一些字还没有进入标准，各单位就把它们放进私用区，再配一款自己的字体来显示。

问题在哪
没有统一含义
：同一个私用区码位，在甲单位的字体里是一个字，在乙单位的字体里是另一个字，在普通字体里什
么都没有；
离开原系统就失效 ：发邮件、导出 PDF
、迁移数据库、换手机，只要对方没有那款专用字体，就显示方框或别的字；
搜索、排序、转换都不认 ：搜索引擎、输入法、简繁转换都不知道这个码位是什么字；
标准已经补上 ：很多当年只能放进私用区的字，后来进入了 Unicode 的正式码位。据 Unicode
汉字专家 Ken Lunde 的介绍， GB 18030-2022 也取消了旧版标准中对私用区映射的要求。

遇到私用区的字怎么办

1. 先查有没有正式码位 ：在 Unihan 数据库中按部首、笔画或字形查找，或看本单位是否有对照表；
2. 有正式码位的，迁移数据 ：把私用区码位批量替换成正式码位，并保留对照表；
3. 确实没有的
：在系统内部继续用私用区过渡，但要建立登记表，同时考虑向标准化机构提交新字申请；对外交换

数据时附上说明，或用汉字描述序列标注，见 《异体字、变体选择符和汉字描述序列》 。

设计字体时的注意
字体设计者有时会在私用区放图标、标志或备用字形。这在字体内部使用没有问题，但不要指望用户在正文里
输入这些码位；需要按特定写法显示的字，优先用标准的变体机制（ OpenType 功能或变体选择符）。

人名、地名里的生僻字：系统怎么做

> 适合谁看：开发会员、订单、政务、金融等要录入人名和地址的系统，不想让生僻字用户「改名」的开发者。

常见的坑

环节	常见问题	正确做法
数据库	MySQL 用 `utf8`（即 `utf8mb3`），存不下 4 字节字符，扩展区的字被截断或报错	库、表、字段、连接全部用 `utf8mb4`
输入校验	用 `[\u4e00-\u9fa5]` 判断「是不是汉字」，把扩展区的字当成非法字符	用 Unicode 属性判断，如 JavaScript 的 `/\p{Script=Han}/u`
长度限制	按 UTF-16 单元计算长度，一个扩展区汉字算 2 个字符，名字被判「太长」	按码位或「字素」计算长度
截断显示	截断时把代理对、变体选择符切开，出现乱码	按码位或字素截断
旧接口	和外部系统用 GBK 交换数据，扩展区的字变问号	用 UTF-8；必须用国标编码时用 GB 18030
显示	字体没有这个字，显示方框	字体栈加补字字体，见《网页和系统里怎么显示生僻字》
打印、证件	打印机或 PDF 没有嵌入字体	生成 PDF 时嵌入有这个字的字体

代码示例

JavaScript 判断是否全是汉字（包括扩展区）：

```
const isHan = (s) => /^\p{Script=Han}+$/u.test(s);
isHan("𠮷"); // true（U+20BB7，扩展 B）
```

按码位计算长度：

```
[..."𠮷"].length; // 2，而 "𠮷".length 是 3
```

MySQL 建表：

```
CREATE TABLE person (
  name VARCHAR(64) CHARACTER SET utf8mb4 COLLATE utf8mb4_0900_ai_ci
);
```

连接时也要指定 `utf8mb4`，否则数据在传输中就会出错。

测试清单

上线前，至少用下面几类字测一遍：

- 扩展 A 的字（基本平面，但很多字体没有）；
- 扩展 B 及以后的字（4 字节）；
- 带变体选择符的字，见《异体字、变体选择符和汉字描述序列》；
- 输入、保存、搜索、导出 Excel、打印、短信、对接第三方，每个环节都要过。

合规要求

用于政务服务和公共服务的产品，GB 18030-2022 要求满足实现级别 3，见《GB 18030：中文编码的国家标准》。

发现缺字怎么反馈

适合谁看：发现某款字体缺了需要的字、字形有错，或者遇到还没有编码的字，想反馈给 “
相关方的人。

先分清是哪一类问题

问题 找谁
某款字体没有这个字，但这个字有 Unicode 码位 字体的作者或厂商
某款字体里这个字写错了、不符合规范 字体的作者或厂商
字有码位，但输入法打不出 输入法厂商
这个字根本没有 Unicode 码位 标准化渠道（见下文）
Unicode 官方资料（如码表、 Unihan 数据）里有错 Unicode 联盟的错误报告页面

反馈时写清楚这几样
1. 哪个字 ：贴出字本身，并写上 Unicode 码位（如 U+20BB7
）；没有码位的，附上清晰图片和汉字描述序列（如 ⿰氵力），写法见
《异体字、变体选择符和汉字描述序列》 ；
2. 用在哪里 ：人名、地名、古籍、专业术语等，附上出处（如证件、地名录、文献页码）；
3. 哪款字体、哪个版本 ：版本号的查法见 《看懂字体信息：名称、版本、字重和许可》 ；
4. 在什么环境下出现 ：操作系统、软件、浏览器；
5. 截图 ：显示问题的截图。

给开源字体反馈
开源字体通常在代码托管平台上有「问题」（ Issues
）页面，按项目说明提交。提交前先搜索有没有人报告过同一个字。

给文风反馈
如果你在文风字体或文风网页字体中发现缺字、字形错误或许可信息有误，请发邮件到
license@wenfeng.org，按上面的要点写清楚。目前请用邮件反馈。

说明：文风收录的字体由各自的作者或厂商制作，缺字能否补上取决于原作者。

没有编码的字
新字进入 Unicode，一般要经过各国家和地区标准化机构汇总后，提交到表意文字工作组（ IRG ）审查，周期以年计。个人可以：

向所在地区的语言文字或标准化主管部门反映；
在此之前，系统内部用私用区过渡，见 《私用区（ PUA）里的字为什么靠不住》 。