《〈原神〉与〈崩坏：星穹铁道〉的文本风格考据》的补充材料

本文件包含：

- 研究语料、研究方法
- 实验设计
- 图 S1 至 S4
- 表 S1 至 S3

1. 研究语料、研究方法

1.1. 研究语料

由于《原神》和《崩坏：星穹铁道》官方并未公开游戏内使用的主线剧情文本[1]，因此我只能从网上收集[2]由网友抄录的文本[1,2]。bilibili wiki（bwiki）使用了自定义的标记语言，需要通过解析网页源代码才可以获得干净的生语料，通过 Python 和正则表达式可以方便地达成这一目的。由于原文本的收集方式为：按照 bwiki 上的分组方式，将源代码分别存储在一个文件中。此外，在研究中，对于那些没有名字的对话，均视为由两个游戏主角所说。

1.2. 研究方法

1.2.1. 语料处理

在 bwiki 中，《原神》的主线剧情文本是按照章节分组的，而《崩坏：星穹铁道》则是按照每章的小节来分组的。这样的分组模式不利于描写两个游戏每一章的统计文体学特征，因此，在获得生语料之后，还需要切分生语料。具体规则为：计算每个文件除标点外的字符数，按每 1000 字为单位，对每个文件实施切分，并分别存储。之后利用自然语言处理（natural language processing, NLP）技术对生语料分词（tokenization）、标注词性（part-of-speech tagging），并做依存句法分析（dependency parsing）以便量化文本的统计学特征，在本研究中，我使用了 HanLP[3]。

1.2.2. 量化

经过 NLP，需要对文本的各项特征进行量化以便开展后续实验。在量化之前，由于切分好的生语料文件可能存在长度不足的情况，我依照生语料长度，剔除了这一部分文件。我结合了不同特征以尽可能实现全面的描写。通过自行编写的 Python 分析脚本和借助外部工具，总共获得了 60 个不同的参数，具体可以分为三部分，分别由基于文章[4] 自行编写的 Python 分析脚本，来自模仿原版 QUITA[5] 编写的用于分析中文文本的 QUITA Python 脚本（除特别标注外，下文中的 QUITA 均指该脚本），外部工具 AlphaReadabilityChinese[3][6] 生成。需要注意，文章[4] 中的“成语使用数”、“成语种类数/成语使用数”、“四字词语使用数”，由于前两个参数较难实现，我选择舍弃这两个参数，并加入对明喻（simile）修辞的计数。同时，由于文章[4] 没有描述“词汇活动度”的具体算法，故该参数由 QUITA 计算。参考文章[7] 加入了“句子节奏度”，参考所述“平均小句长”的算法。参考文章[8]，加入了平均依存距离（Mean Dependency Distance, MDD），在实际应用过程中，由于每个文件都是由若干个句子组成，MDD 是对每个单独的句子作计算，因而为了表示整个文件，我对每个文件中计算得到的 MDD 再次求平均，并求出相应的标准差，两个参数均记录在生成的参数文件中。此外，原版 QUITA 中提供了类符/形符比（Type-Token Ratio, TTR），但由于 TTR 易受文本长度影响[9]，所以在这个研究中，我使用了标准化类符/形符比（Standardized Type-Token Ratio, STTR），该参数按照将 200 词（我使用的是按照词为切分单位，而不是字）归入一组，分别计算 TTR（如果最后一组词数量不足 200，但不少于 200 词的 90%（180 词），则视为 200 词，并参与计算），经过计算，获得到一组 TTR，最后求该组 TTR 值的平均值，即为 STTR。

量化完成之后还需要标注样本，即说明样本属于哪个游戏，来自哪个游戏的哪个章节。标注规则如下：（1）对于文本来自哪个游戏的标注主要采用二分法，即对于来自《崩坏：星穹铁道》的文本都标注为 0，来自《原神》的文本均标注为 1。（2）对于章节的标注则采用游戏内的章节号，需要注意的是，序章均标注为 0，《原神》的间章则从 90 开始编号。

---

[1] 主线剧情指《原神》的“魔神任务”、《崩坏：星穹铁道》的“开拓任务”

[2] 《原神》主线剧情文本的访问日期均为：2024 年 4 月 4 日；《崩坏：星穹铁道》主线剧情文本的访问分多次完成，具体时间见表 S 3。

[3] https://github.com/leileibama/AlphaReadabilityChinese

1.2.3. 特征识别

为了更好地发现文本的计量特征，这个研究中我使用了机器学习来学习特征、识别特征。这里有一个非常重要的假设——如果模型由良好的表现，那么可以认为不同文本之间的特征是明显的。本研究中使用了 9 种不同模型，分别是最近邻模型、线性支持向量机（Support Vector Machine, SVM）、径向基 SVM、决策树、随机森林、神经网络、AdaBoost、朴素贝叶斯分类器、极端梯度提升（eXtreme Gredient Boost，XGB），在其他实验中还使用了线性判别分析、层次聚类。在实验时，我发现不同游戏、不同章节间，文本的数量存在很大差异，这就造成了数据集的不平衡。不平衡数据集会影响模型表现，导致模型可能会错误分类样本较少的数据[10]。鉴于此问题在这个研究中也出现了，我使用了 SMOTE 算法[10] 来平衡不同样本间的数据，以提升模型的表现。

1.2.4. 性能度量

我采用多种指标（表 S 1），全面衡量模型表现。在训练和测试模型时，对于每个模型，我都是从平衡过的数据中随机抽取 70% 的样本作为训练集，余下 30% 做测试集，如此重复 100 次，得到每个模型的各项指标（表 S 1）。

表 S 1|本研究中用于衡量模型学习效果的指标、算法。算法中文名、具体算法均来自《机器学习》[11]29–35。AUC 和 Acc 公式中每一项符号的具体含义见下文所述。

名称	缩写	算法
曲线下面积（Area Under Curve）	AUC	$\frac{1}{2}\sum_{i=1}^{n-1}(x_{i+1}-x_i)\times(y_i+y_{i+1})$
准确度（Accuracy）	Acc	$\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x_i})\neq y_i)$
查准率（Precision）	Pre	$\frac{\text{真正例数}}{\text{真正例数}+\text{假正例数}}$
查全率（Recall）	Rec	$\frac{\text{真正例数}}{\text{真正例数}+\text{假反例数}}$
F1-度量（F1-Score）	F1S	$\frac{2\times\text{查准率}\times\text{查全率}}{\text{查准率}+\text{查全率}}$

接下来我会逐个介绍每一个指标的具体含义，但是考虑到连贯性，我将打乱表格的顺序。首先会介绍混淆矩阵，然后分别是查准率、查全率，接着是 AUC 的基础 ROC（Received Operating Characteristic，受试者工作特征）曲线，之后会介绍 AUC 的含义，其后是介绍 F1-度量，最后介绍准确度（因为准确度用了不同的算法，所以与上边几项无关，因而放在最开始和最后讲都可以）。本研究所用性能度量（表 S 1）的介绍，除特殊标注外，均来自《机器学习》[11]29–35。

1.2.4.1. 混淆矩阵

混淆矩阵（Confusion Matrix，表 S 2）是理解表 S 1 中除准确度之外其它各项指标的基础，在此做简要介绍。机器分类结果可以分为四类，分别是真正例、假正例、假反例、真反例。举例说明以上四种情形，这里采用文章[12] 中的例子来说明。假设现在需要机器学会根据肿瘤标志（Tumor Marker，TM）的数值分辨那些肿瘤是恶性的，那些是良性的。那么，真正例就是

指根据 TM 机器正确地将恶性肿瘤的样本正确分类到了恶性肿瘤这个集合中；真反例是指机器将良性肿瘤的样本正确分类到了良性肿瘤这个集合中。假正例是指，机器将良性肿瘤误判为恶性肿瘤；假反例是指机器将恶性肿瘤分类为了良性肿瘤。显然，只有主对角线（从左上到右下）上的预测结果（真正例、真反例）是正确的预测结果。副对角线上的假反例和假正例都属于误判。

表 S 2|分类结果混淆矩阵

实际情况	预测结果	
	正例	反例
正例	真正例	假反例
反例	假正例	真反例

1.2.4.2. 查准率、查全率

根据查准率公式（表 S 1），它代表了在分类为正例的样本中，真正是正例的样本占所有机器分类为正例的结果的比例。这个数值越高，代表着机器的正确分类能力越好。查全率代表了在所有正例样本中，有多少样本被机器正确分类为正例。

查全率与查准率是一对相互对立的度量。原因在于，不同的任务对查准率和查全率的要求不同，如果我们更多要求机器能准确分类，那么要求机器只挑选那些最有把握的样本作为正例，但是可能也会有很多样本没能正确分类，这就导致了查全率的下降。相反，如果我们需要尽可能多的筛选出来正例，那么一些反例就有可能被误选进来，这样就使查准率下降了。在一些简单任务中，二者都有可能会很高。

对于本研究来说，查准率代表了机器从不同文本中识别出某一类文本（比如某一章节或某一游戏的文本）的能力，查全率则表示对同一类型的文本，机器能从中分辨出多少文本是来自这一类型。

1.2.4.3. 曲线下面积

在计算 AUC 之前，需要先获得 ROC（Received Operating Characteristic，受试者工作特征）曲线，AUC 中的“面积”就是指 ROC 曲线下面积。机器在分类正例和反例的时候是依据某个截断点（cut point），如果输出值大于截断点则判为正例，否则为反例。那么，如果让截断点的值从小到大变化，在每次变化中都会产生一个混淆矩阵，根据此表，结合以下两个公式：

$$\begin{aligned} 真正例率 &= \frac{真正例数}{真正例数 + 假正例数} \\ &= 查准率, \end{aligned}$$

$$假正例率 = \frac{假正例数}{假正例数 + 真反例数},$$

便可绘制出一条 ROC 曲线[12]。注意，“假正例率”沿用了《机器学习》[11]34的说法。在文献[12]中，“真正例率”一般称为 Sensitivity（敏感度），“假正例率”则是用 1 减去 Specificity（特异度）的差值。

在实际情况中，ROC 曲线的绘制都是由有限个样本完成的，因而曲线不是连续的。假设某 ROC 曲线由一系列坐标分别为$\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$的点构成，那么利用积分知识便可得到表 S 1 中的 AUC。

1.2.4.4. F1-度量

F1-度量是查准率和查全率的调和平均，即

$$\frac{1}{\mathrm{F1}}=\frac{1}{2}\times\left(\frac{1}{\text{查准率}}+\frac{1}{\text{查全率}}\right),$$

它基本的形式为：

$$F=\frac{(\beta^2+1.0)\times\text{查准率}\times\text{查全率}}{\beta^2\times\text{查准率}+\text{查全率}}。$$

如果$\beta=1$（F1-度量）则说明查准率和查全率同等重要[13]。

1.2.4.5. 准确率

机器在判别一个样本是正例还是反例的时候，存在误判的可能性，也就是说，把原本的正例误判成反例，或者反过来。准确率就是衡量在所有样本中，有多少样本被正确分类了。表 S 1 中的指示函数Ⅱ(•)是指，对于给定样本$\boldsymbol{x_1}$，机器学习算法$f(\cdot)$的输出值如果与样本原本标签y_1一致，即$f(\boldsymbol{x_1})=y_1$，则指示函数输出值为 1，否则为 0。遍历所有样本（假设有m个样本），对指示函数求和，再除以样本数，便是准确率。准确率度量一个算法正确分类的能力。查准率和查全率与准确率的侧重有所不同，前两个更加强调机器识别正例的能力，而后者（准确率）则全面考虑了机器识别正确正例和反例的能力[14]。

1.2.5. 模型输出解释

机器学习已经在自然科学领域得到了广泛应用，但是由于其结构往往非常复杂，也正是由于这种复杂性让机器的学习能力不断提升、准确率不断上升[15]21。同时，理解机器得出某个结果（prediction）是使机器学习能用于研究领域的前提[16]。因此，为了能解释机器的结论，我用到了夏普利加性解释（SHapley Additive exPlanations，SHAP）[17]。

SHAP 的主要思想就是将机器学习模型视为一个“黑盒”（black box），观察特征对模型输出值的影响，具体公式如下：

$$\phi_{i(f,x)}=\sum_{R\in\mathfrak{R}}\frac{1}{M!}\left[f_x\left(P_i^R\cup i\right)-f_x\left(P_i^R\right)\right],$$

其中，$\mathfrak{R}$是样本全部特征的集合，P_i^R在特征i之前的所有特征，M是样本特征数量[18]，f_x则是模型。其中的ϕ_i就是一个 SHAP 值[17]。从这个公式中可以看出，为了计算 SHAP 值需要两个模型，第一个模型的输入是特征i以及之前的所有特征，第二个模型的输入只有特征i之前的所有特征，这样就可以计算出特征i对模型输出的贡献。

根据模型筛选的结果，由于 XGB 表现出众，我主要是使用了属于 SHAP 之一的 TreeSHAP[18]。同时，相较其他算法，SHAP 的优势[18] 有：（1）独立于树深度的变化，公正地给样本的每个特征（feature）分配重要性；（2）结果不会发生改变；（3）SHAP 的结果更加符合人类直觉。

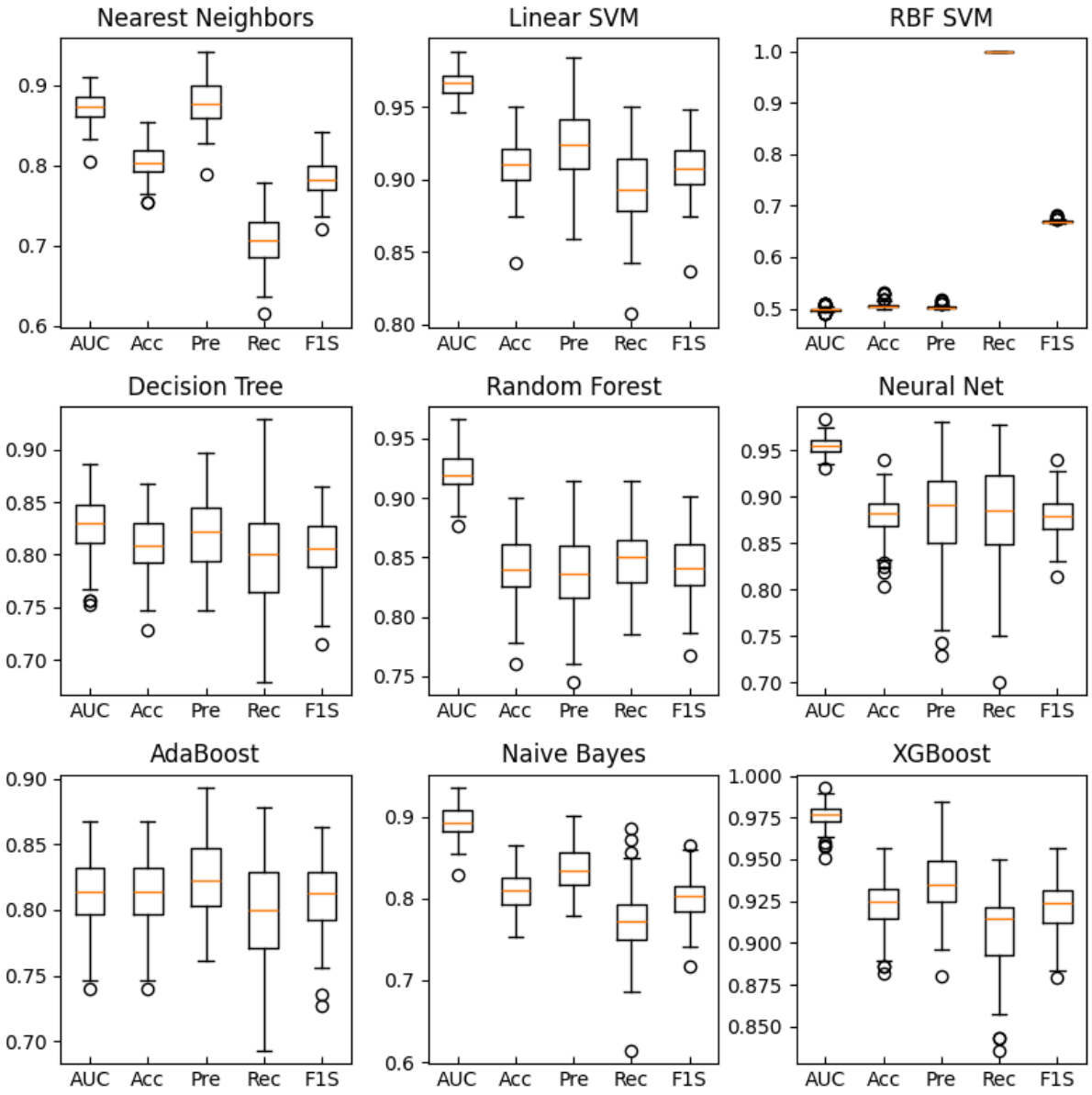

图 S 1|实验一中用到的机器学习模型及其表现。方框的上部为上四分位（75%），下部为下四分位（25%）；方框中间的橙色实线是中位数，圆圈为异常值。

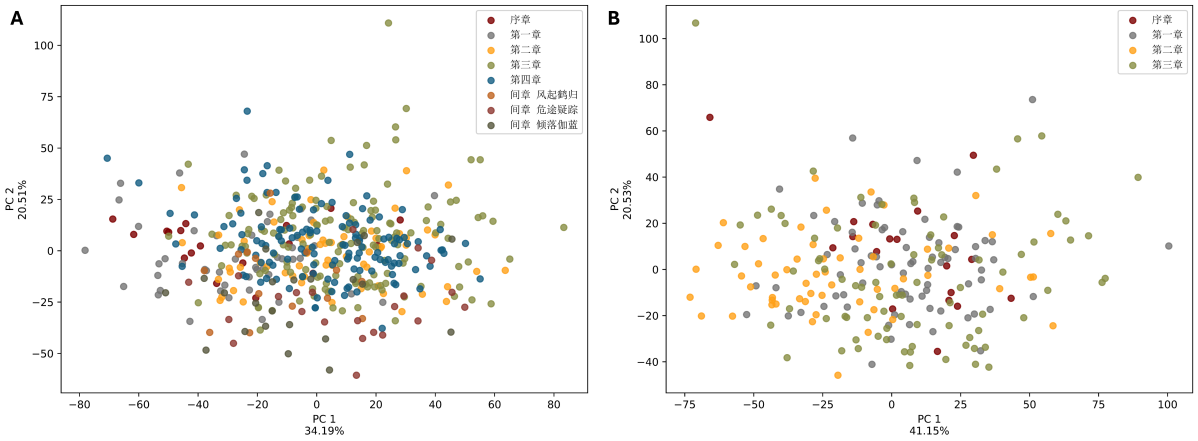

图 S 2|利用 PCA 可视化的两个游戏不同章节文体计量特征。**A** 为《原神》，**B** 为《崩坏：星穹铁道》。PC1、PC2 分表代表两个主成分，下边的数字为该主成分解释的方差，即包含了的原始信息的量。

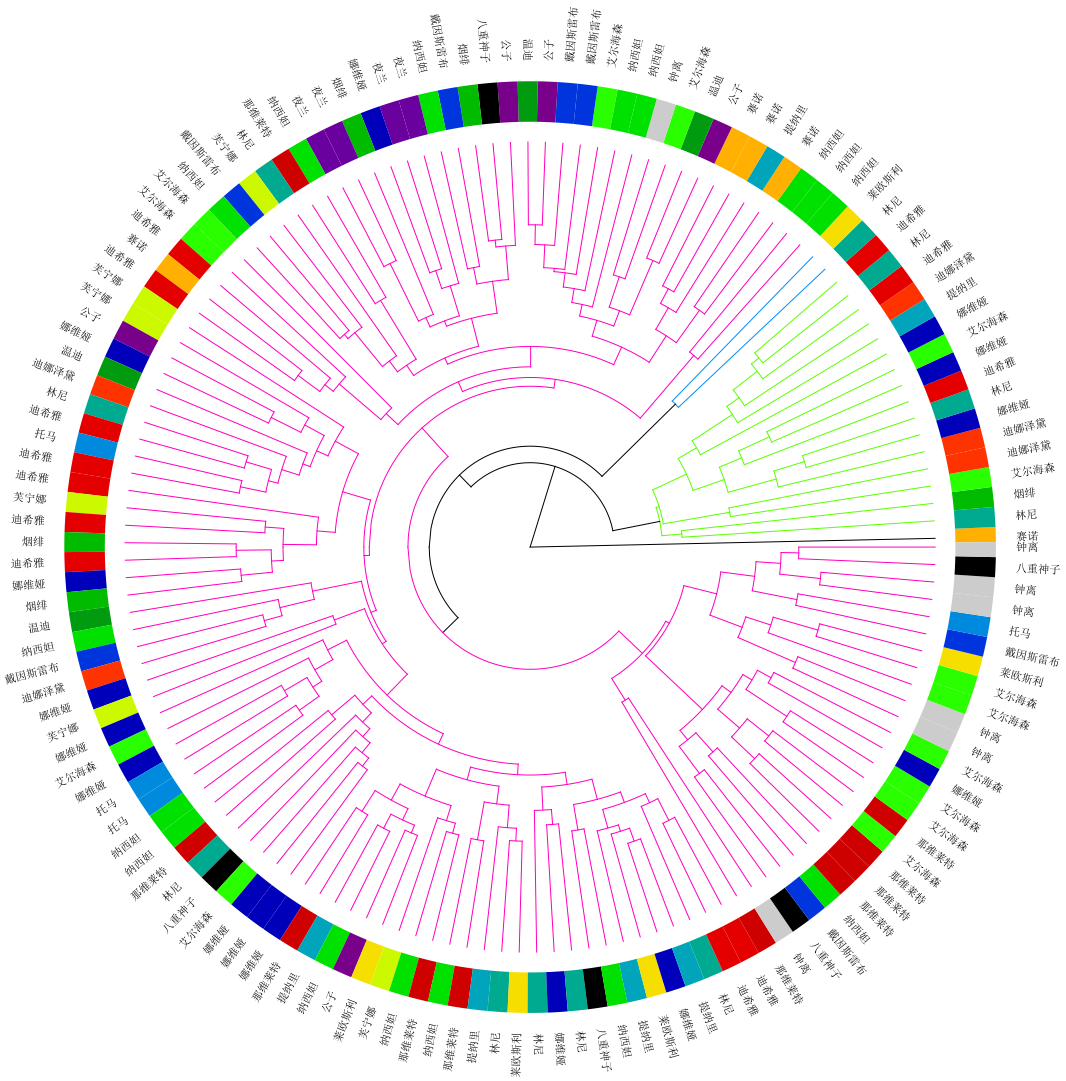

图 S 3|《原神》中不同角色的对话文本的特征聚类。其中去除了对话文本数量较多的旅行者、派蒙。

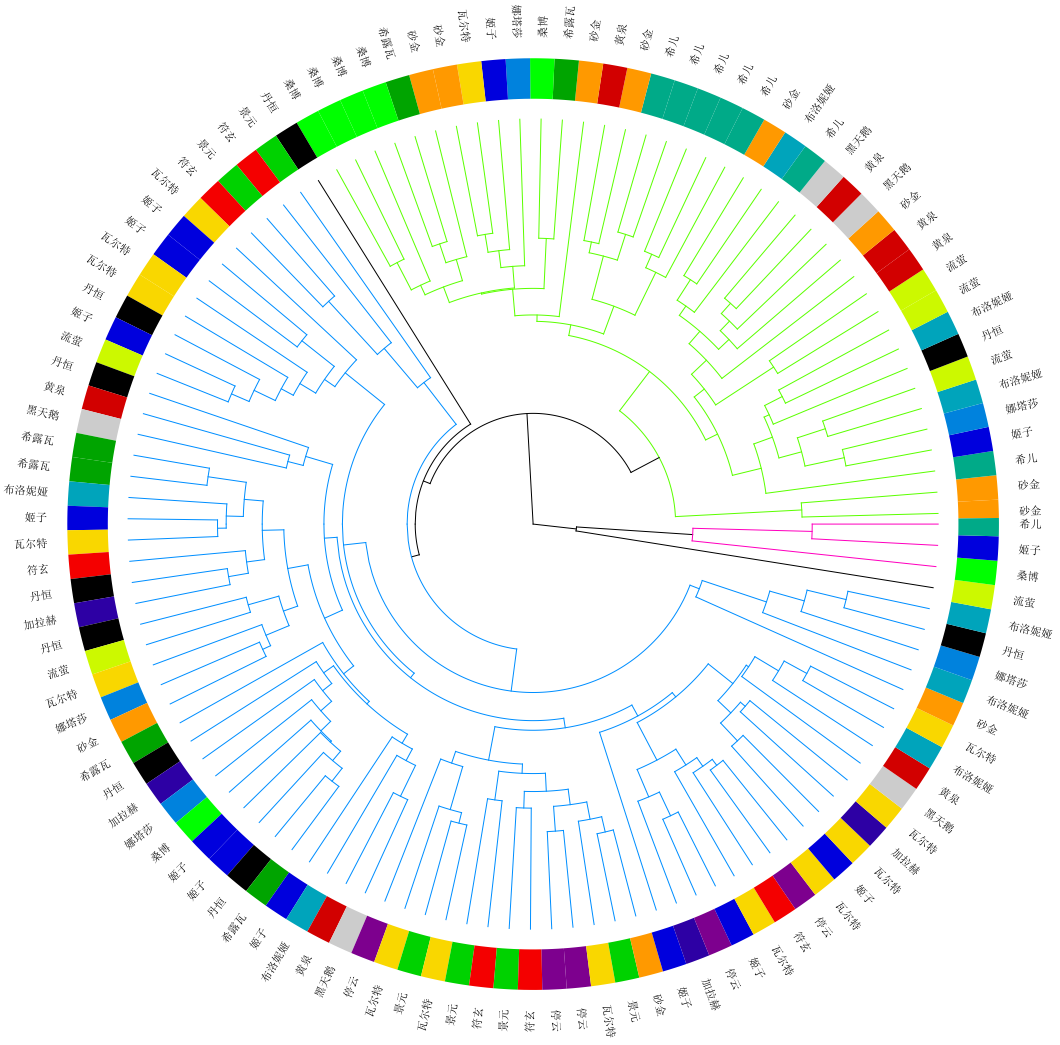

图 S 4|《崩坏：星穹铁道》中不同角色的对话文本的特征聚类。其中去除了对话文本数量较多的三月七、开拓者。

表 S 3|《崩坏：星穹铁道》主线剧情各个小节访问时间。在文本收集过程中，由于我是直接复制网页的源代码， 所以文件的创建时间可以认为是网页访问时间。

访问时间	章节
2024-04-04	1.2.10_时不我待，我的朋友、 1.2.11_静静的星河、 1.2.1_在屋外的黑暗中洗涤、 1.2.2_不可制造偶像、 1.2.3_青年近卫军、 1.2.4_兵士们默默无言、 1.2.5_星星是冰冷的玩具、 1.2.7_这里的黎明……、 1.2.8_回归、 1.2.9_从凶险和泥泞的沼泽中、 2.1.10_诸天无安，迷途难返、 2.1.11_茸客鸣呦，玉角盘虬、 2.1.1_旅进青霄，不速之邀、 2.1.2_行遏流云，身入魔阴、 2.1.3_紫府通谒，将军定策、 2.1.4_旧影婆娑，追思错落、 2.1.5_犬迹追从，谛听狐踪、 2.1.6_迴星周旋，未卜知先、 2.1.7_长乐新朋，青鸟候风、 2.1.8_极数问玄，历事穷观、 2.1.9_神木重萌，掣转天衡、 2.2.1_金鼎灵树，穷途椅杌、 2.2.2_螣蛇无穴，旧梦亡阙、 2.2.3_得其雨露，安其壤土、 2.2.4_有龙矫矫，其渊渺渺、 2.2.5_仙骸成空，大劫有终、 2.3.1_安灵布奠，天清路远、 3.1.10_倘若在午夜醒来、 3.1.11_是谁杀死了…、 3.1.1_长日入夜行、 3.1.2_丑时三刻的敲门声、 3.1.3_那些逐梦的年轻人、 3.1.4_无眠之夜、 3.1.5_黄金年代的故事、 3.1.6_好兆头，我的朋友、 3.1.7_北风的安眠曲、 3.1.8_夜色名为温柔、 3.1.9_犹在镜中
2024-04-05	0.1.1_混乱行至深处、 0.1.2_漩涡止于中心、 0.1.3_宇宙安宁片刻、 0.1.4_模拟宇宙-始发测试、 0.1.4_阴影从未离去、 0.1.5_旅途正在继续、 0.1.6_星间流浪、 1.1.10_已故去的必如雪崩再来、 1.1.11_躺在铁锈中、 1.1.12_腐烂或燃烧、 1.1.13_我们不擅长告别、 1.1.1_激「冻」人心的大冒险、 1.1.2_如果在冬夜，一群旅人、 1.1.3_永冬城之夜、 1.1.4_躲得过初一，躲不过十五、 1.1.5_捉迷藏、 1.1.6_第八条、也是最后一条规则、 1.1.7_她等待刀尖已经太久、 1.1.8_他们有多少人已掉进深渊、 1.1.9_相会在日落时分、 3.2.1_天鹅绒里的恶魔、 3.2.2_迷惘的一代人、 3.2.3_双重赔偿、 3.2.4_酒店关门之后、 3.2.5_外邦为何争闹?、 3.2.6_人间天堂、 3.2.7_泄密的心
2024-04-06	3.2.8_所有悲伤的故事、 3.2.9_行过死荫之地

表 S 4|各模型依据文本计量特征识别《原神》不同章节的性能度量。

		AUC	Acc	Pre	Rec	F1S
Nearest Neighbors	**Median**	0.9329	0.7926	0.7782	0.7926	0.7693
	STD	0.0072	0.0151	0.0198	0.0151	0.0177
Linear SVM	**Median**	0.9738	0.8452	0.8402	0.8452	0.8401
	STD	0.0043	0.0148	0.0163	0.0148	0.0156
RBF SVM	**Median**	0.5204	0.1705	0.7664	0.1705	0.1137
	STD	0.095	0.0119	0.1211	0.0119	0.0212
Decision Tree	**Median**	0.8671	0.6051	0.6223	0.6051	0.5999
	STD	0.0166	0.0359	0.0321	0.0359	0.0365
Random Forest	**Median**	0.9566	0.7756	0.7709	0.7756	0.7655
	STD	0.0056	0.0221	0.0235	0.0221	0.0231
Neural Net	**Median**	0.9496	0.7116	0.7435	0.7116	0.7019
	STD	0.0069	0.0352	0.0246	0.0352	0.0374
AdaBoost	**Median**	0.8377	0.7159	0.7103	0.7159	0.7104
	STD	0.0147	0.0257	0.0265	0.0257	0.0259
Naive Bayes	**Median**	0.9467	0.7045	0.7065	0.7045	0.7001
	STD	0.0066	0.022	0.0228	0.022	0.0227
XGBoost	**Median**	0.9838	0.8693	0.8629	0.8693	0.8634
	STD	0.0033	0.0144	0.0157	0.0144	0.0147
LDA	**Median**	0.9779	0.8438	0.8382	0.8438	0.8392
	STD	0.0036	0.0155	0.0167	0.0155	0.0161

表 S 5|各模型依据文本计量特征识别《崩坏：星穹铁道》不同章节的性能度量。

		AUC	Acc	Pre	Rec	F1S
Nearest Neighbors	**Median**	0.8446	0.6458	0.6559	0.6458	0.6261
	STD	0.0228	0.0385	0.044	0.0385	0.0434
Linear SVM	**Median**	0.9049	0.7292	0.7406	0.7292	0.7267
	STD	0.0224	0.0384	0.039	0.0384	0.0391
RBF SVM	**Median**	0.4896	0.2812	0.5638	0.2812	0.1581
	STD	0.0543	0.0312	0.201	0.0312	0.0509
Decision Tree	**Median**	0.7526	0.5938	0.5972	0.5938	0.5922
	STD	0.034	0.0458	0.0472	0.0458	0.0461
Random Forest	**Median**	0.906	0.7083	0.7126	0.7083	0.7017
	STD	0.0185	0.0404	0.0433	0.0404	0.042
Neural Net	**Median**	0.8976	0.7083	0.726	0.7083	0.7057
	STD	0.021	0.0446	0.039	0.0446	0.0441
AdaBoost	**Median**	0.7269	0.6146	0.6143	0.6146	0.6081
	STD	0.0358	0.0479	0.052	0.0479	0.0497
Naive Bayes	**Median**	0.8867	0.6771	0.6874	0.6771	0.6777
	STD	0.0233	0.0411	0.0399	0.0411	0.0415
XGBoost	**Median**	0.9455	0.7917	0.7898	0.7917	0.7851
	STD	0.0158	0.0389	0.0411	0.0389	0.0403
LDA	**Median**	0.9324	0.7812	0.785	0.7812	0.7758
	STD	0.0187	0.0369	0.0355	0.0369	0.038

参考文献

[1] 原神 WIKI 贡献者. 《原神》魔神任务[EB/OL]. https://wiki.biligame.com/ys/%E9%AD%94%E7%A5%9E%E4%BB%BB%E5%8A%A1.

[2] 崩坏：星穹铁道 WIKI 贡献者. 《崩坏：星穹铁道》开拓任务[EB/OL]. https://wiki.biligame.com/sr/%E5%BC%80%E6%8B%93%E4%BB%BB%E5%8A%A1.

[3] HE H, CHOI J D. The Stem Cell Hypothesis: Dilemma behind Multi-Task Learning with Transformer Encoders[C/OL]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 5555-5577. https://aclanthology.org/2021.emnlp-main.451.

[4] 仲文明, 姚梦妮. 基于降维分类模型的译者风格研究——以 Silent Spring 五译本为案例[J]. 外语电化教学, 2023, 0: 24-31.

[5] KUBÁT M, MATLACH V, ČECH R. QUITA – Quantitative Index Text Analyzer[M]. 2014.

[6] 雷蕾, 韦瑶瑜, 刘康龙. AlphaReadabilityChinese：汉语文本可读性工具开发与应用[J]. 外语与外语教学, 2024, 0: 83-93.

[7] 仲文明, 王靖涵. 少年儿童翻译文学的译本风格计量研究——以 Silent Spring 三译本为例[J]. 外语与翻译, 2023, 30: 20-27.

[8] 刘海涛. 基于依存树库的汉语句法计量研究[J]. 长江学术, 2008, 0: 120-128.

[9] CVRČEK V, CHLUMSKÁ L. Simplification in translated Czech: a new approach to type-token ratio/Упрощение в чешских переводных текстах: новый подход к отношению словоформа/словоупотребление (type-token ratio)[J]. Russian Linguistics, 2015: 309-325.

[10] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[11] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[12] NAHM F S. Receiver operating characteristic curve: overview and practical use for clinicians[J]. Korean journal of anesthesiology, 2022, 75(1): 25-36.

[13] CHINCHOR N. MUC-4 evaluation metrics[C/OL]//Proceedings of the 4th Conference on Message Understanding. McLean, Virginia: Association for Computational Linguistics, 1992: 22-29. https://doi.org/10.3115/1072064.1072067. DOI:10.3115/1072064.1072067.

[14] POWERS D M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation[J]. arXiv preprint arXiv:2010.16061, 2020.

[15] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep Learning[M]. MIT Press, 2016.

[16] ROSCHER R, BOHN B, DUARTE M F, et al. Explainable machine learning for scientific insights and discoveries[J]. Ieee Access, 2020, 8: 42200-42216.

[17] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017: 4768-4777.

[18] LUNDBERG S M, ERION G, CHEN H, et al. From local explanations to global understanding with explainable AI for trees[J]. Nature machine intelligence, 2020, 2(1): 56-67.