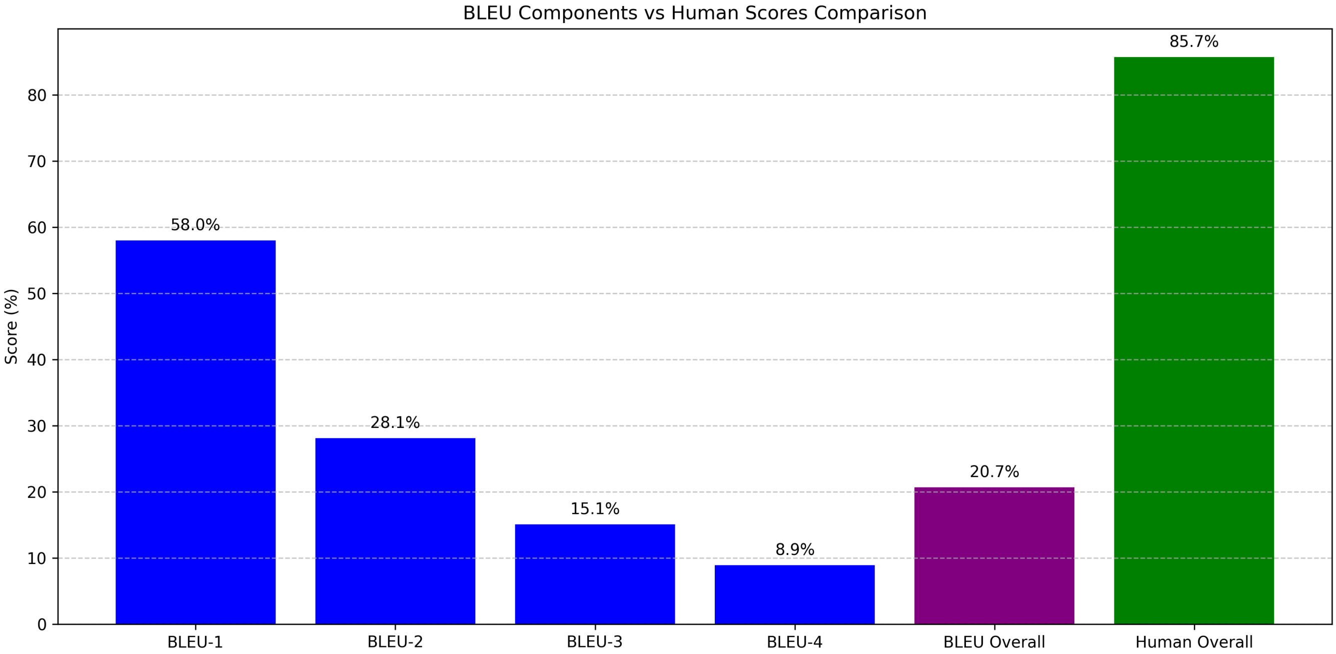
Appendix B: BLEU Evaluation Details

B.1 BLEU Score Details

Our BLEU evaluation yielded the following results:

```
{
 "name": "BLEU",
 "score": 20.7,
 "signature": "nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.5.1",
 "verbose_score": "58.0/28.1/15.1/8.9 (BP = 0.955 ratio = 0.956 hyp_len =
8755 ref_len = 9155)"
}
```



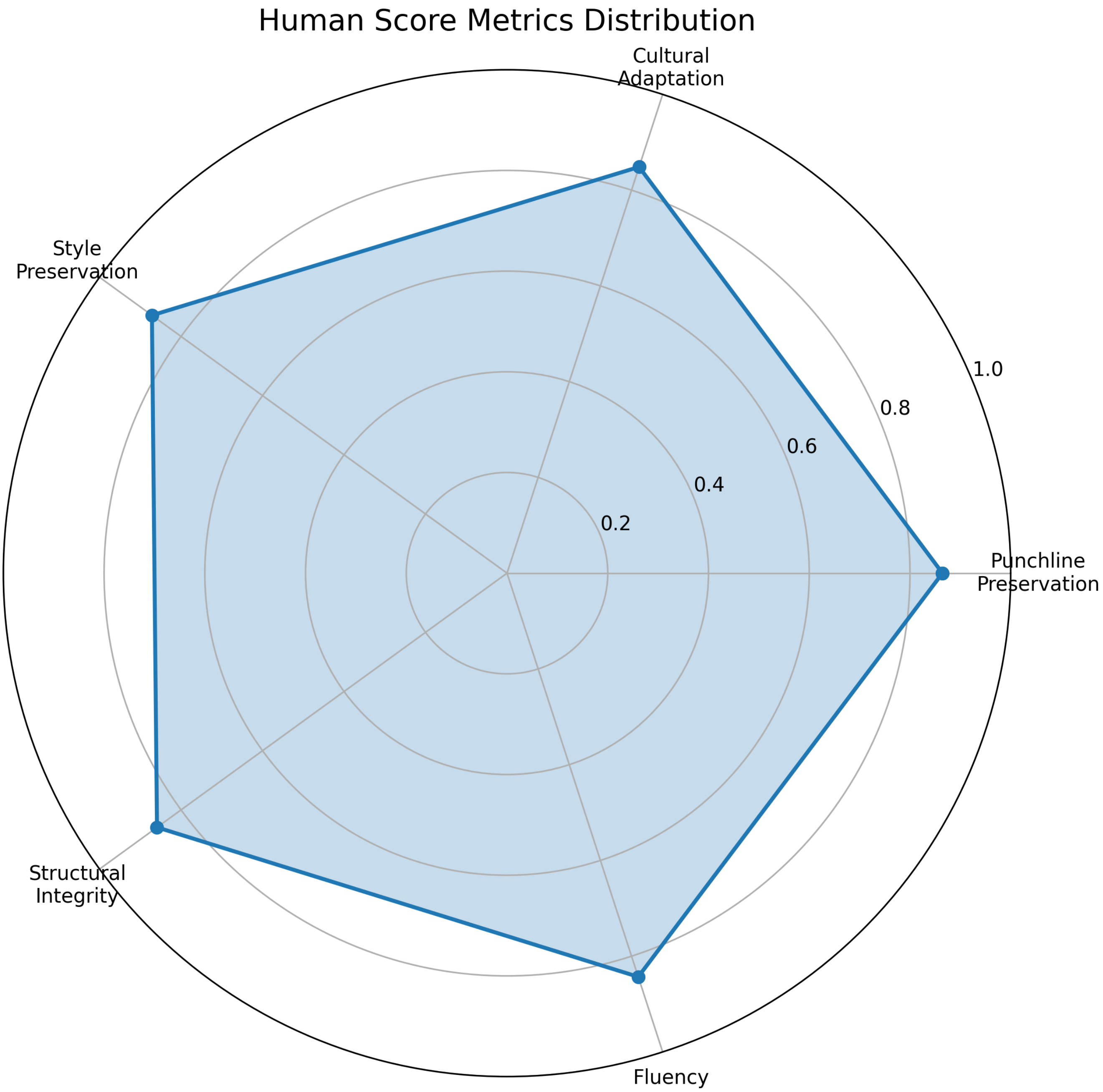
The BLEU score of 20.7 can be broken down into its n-gram precision components:

- **Unigram precision (BLEU-1)**: 58.0%
- **Bigram precision (BLEU-2)**: 28.1%
- **Trigram precision (BLEU-3)**: 15.1%
- **4-gram precision (BLEU-4)**: 8.9%
- **Brevity penalty (BP)**: 0.955 (translation length is 95.6% of reference length) We also calculated the chrF2 score:

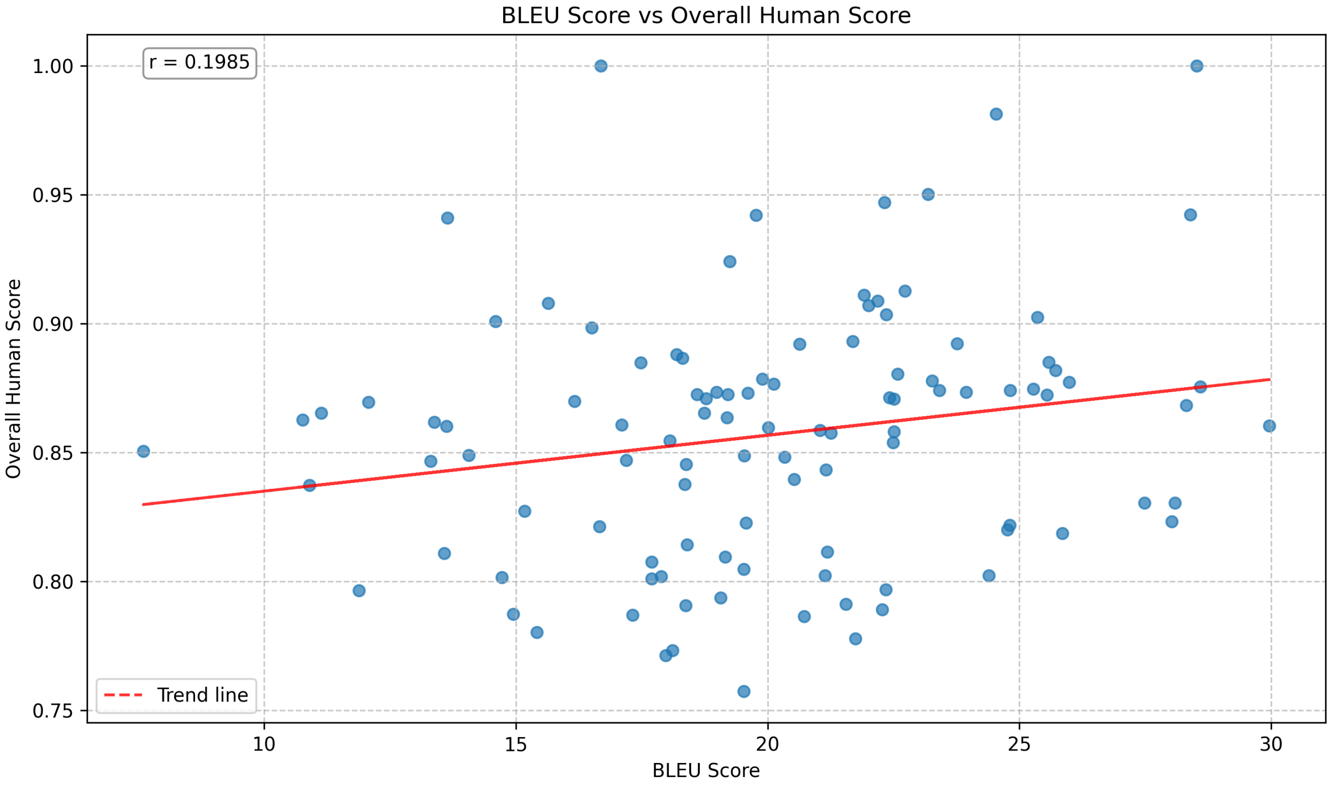
```
{
 "name": "chrF2",
 "score": 19.4,
 "signature":
"nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1"
}
```

Custom score averages from our human evaluation:

- **Overall quality**: 0.8572
- **Punchline preservation**: 0.8645
- **Cultural adaptation**: 0.8485
- **Comedian style preservation**: 0.8718
- **Structural integrity**: 0.8590
- **Fluency**: 0.8431



B.2 Correlation Between BLEU and Custom Scores



- **BLEU vs overall**: 0.2219
- **BLEU vs punchline preservation**: 0.2423
- **BLEU vs cultural adaptation**: 0.2121
- **BLEU vs comedian style preservation**: 0.1196
- **BLEU vs structural integrity**: 0.2204
- **BLEU vs fluency**: 0.1933

B.3 Analysis of BLEU Score Limitations for Comedy Translation

While our BLEU score of 20.7 would be considered moderate in general machine translation evaluation, our analysis reveals several key limitations when applying BLEU to comedy translation:

1. Misalignment with creative translation goals: The core objective of stand-up comedy translation is to preserve humor and cultural connotations, not achieve word-for-word correspondence. While our BLEU-1 score of 58.0% shows reasonable word-level matching, the rapidly declining n-gram scores (down to 8.9% for BLEU-4) highlight how creative rewriting diverges from literal translation.

2. Inability to capture punchline effectiveness: Despite achieving high human scores for punchline preservation (0.8645), the correlation between BLEU and this critical dimension is only 0.2423. This demonstrates BLEU's fundamental inability to evaluate whether the humor effect is preserved.

3. Penalization of cultural adaptation: Our translations deliberately adapt cultural references to resonate with Chinese audiences. While this approach receives high human scores for cultural adaptation (0.8485), it reduces n-gram matches with reference translations, explaining the weak correlation of 0.2121.

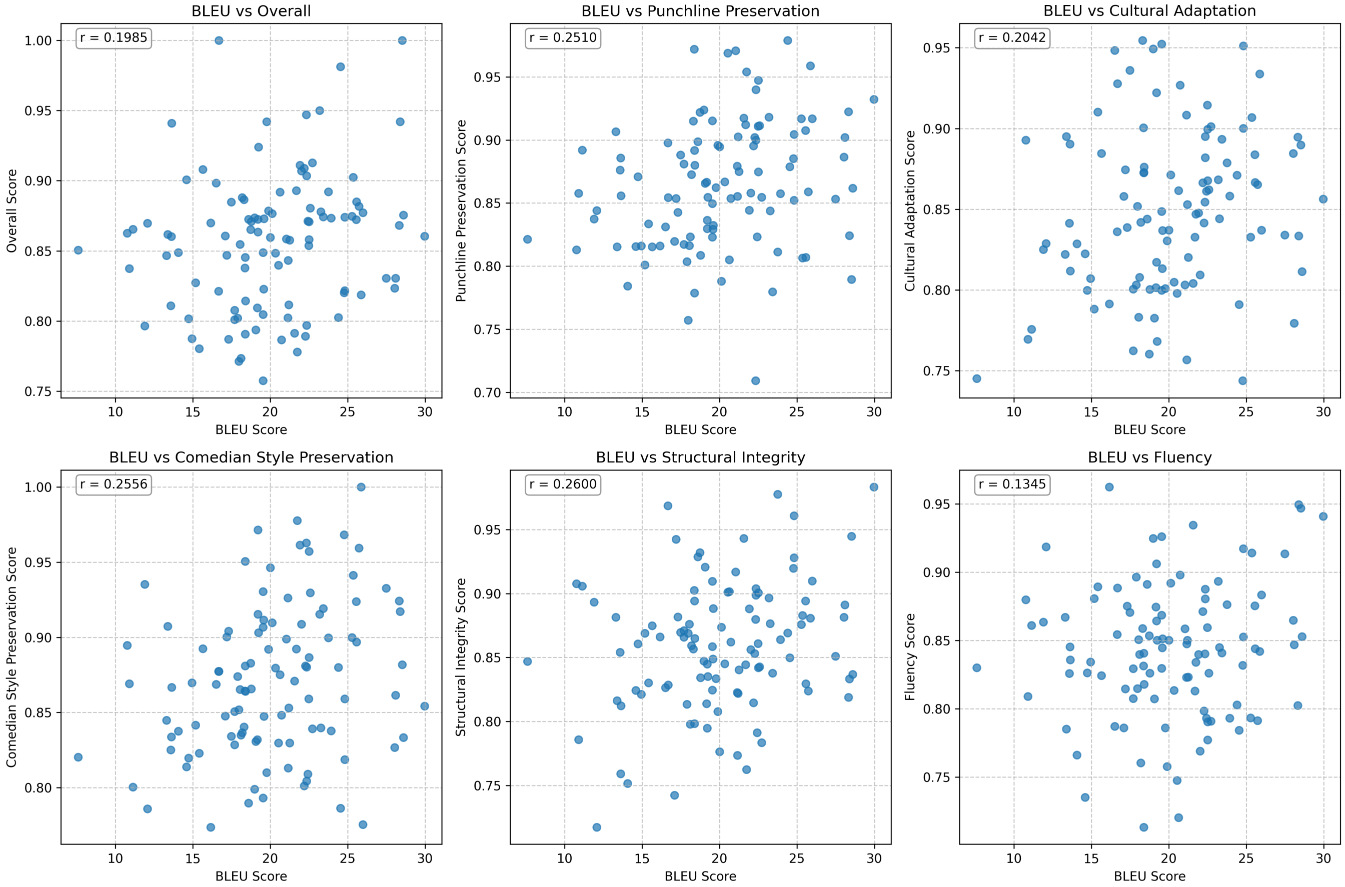
4. Limited recognition of stylistic choices: The lowest correlation (0.1196) is between BLEU and comedian style preservation, confirming that BLEU cannot recognize when a translation successfully captures a comedian's unique voice and delivery style.

5. Surface-level evaluation: The moderate BLEU-1 score (58.0%) compared to the low BLEU-4 score (8.9%) indicates that while individual words may be preserved, the creative restructuring of phrases and sentences—essential for effective comedy translation—is penalized by BLEU.

6. Brevity penalty limitations: The brevity penalty of 0.955 slightly reduces our BLEU score, yet our translations are deliberately crafted to maintain optimal length for comedic timing and audience comprehension. These findings do not indicate poor translation quality; rather, they reveal the fundamental mismatch between traditional evaluation methods like BLEU and the requirements of creative translation. An excellent stand-up comedy translation with human scores averaging 0.85+ might receive a moderate BLEU score around 20, demonstrating the need for a specialized evaluation framework for creative content translation.

B.4 Visualization

The following visualizations illustrate the relationships and findings discussed:



These images provide a visual representation of the correlations and limitations of the BLEU score in evaluating comedy translation quality.