Translating Bridge Game Records into English Through a Controlled Artificial Language

Kevin Chen
New York University
Natural Language Processing
kevin734626@gmail.com

Abstract

Bridge game records encode auctions, contracts, opening leads and play in a compact notation that beginners cannot read. I treat this notation as a source language and ask whether translating it first into a controlled artificial language (AL), a predicate-style meaning representation that labels each call with its role and conventional meaning, helps a language model explain the record in English. I implement a rule-based parser, the AL, a legality checker, template generators and baselines, and compare two local language models (Qwen3-4B-Instruct and Qwen3-1.7B) given either the raw record or its AL translation with identical instructions. The main test set is 40 real boards from Bridge Base Online vugraph broadcasts, chosen so that no auction pattern appears in the training or development data. Given the AL instead of the raw record, the 4B model explains 63% of annotated bid meanings instead of 21% (paired bootstrap 95% CI for the gain: $+0.31$ to $+0.53$), with unchanged fact accuracy (0.90). The 1.7B model improves from 5% to 31%. A template generator reading the same AL states every fact and explains 64% of meanings, statistically indistinguishable from the best model. Error analysis shows that the remaining gap is mostly knowledge the AL does not contain: conventions such as Multi 2$\diamondsuit$ and the weak notrump that the hand-written rules do not model.

1 Introduction

Bridge is a card game with a highly compressed notation. A short sequence such as `North 1NT -- East Pass -- South 3NT -- All Pass` tells an experienced player the shape and strength of North's hand, that South judged the partnership to have enough for game, and who will play the hand. A beginner sees letters and numbers. Explaining the record requires knowledge that is not written down: what a bid conventionally shows, which partnership a player belongs to, whether a double asks partner to bid or to defend, and how all of this depends on what was bid before.

This makes Bridge commentary a data-to-text generation problem (Reiter and Dale, 2000) with an unusual property. The input is fully structured, but most of its meaning is implicit and context-dependent. The same call, `Double`, asks partner to bid a suit in one auction and to defend for penalty in another. A system that verbalizes the input faithfully can still explain nothing.

I study whether an explicit intermediate representation helps. The research question is:

> Does translating a Bridge record into a controlled artificial language before generation improve the factual accuracy and the explanation of bid meanings in English produced by a language model?

The AL is produced by a rule-based parser and makes partnership roles and conventional meanings explicit, for example `OVERCALL(West, 2H, artificial)` or `DOUBLE(South, takeout)`. The experiment holds the language model and its instructions fixed and changes only the input: raw record or AL.

The contributions are:

1. A parser and AL for contested Bridge auctions, with role-aware labels (opening, response, overcall, advance, rebid), seat-aware openings, and context-dependent classification of doubles and redoubles, plus a legality checker for auctions and play.
2. A real-data test set of 40 boards from top-level team events, selected so that every auction pattern is unseen during development, with full hands, verified results, reference explanations and 105 annotated bid meanings.
3. A controlled comparison of two local Qwen3 models on raw versus AL input, scored

with role-aware fact accuracy, meaning recall, a hallucinated-bid count, readability and reference-based metrics, with paired bootstrap confidence intervals.

The main result is that the AL roughly triples the share of bid meanings the models explain without costing facts. The model adds little knowledge of its own: it mostly verbalizes what the AL already says, so the quality of the explanation is limited by the quality of the representation.

2 Related Work

Data-to-text generation. Generating text from structured records has a long history, from pipeline systems with explicit content selection and microplanning (Reiter and Dale, 2000) to neural models. Datasets such as WebNLG (Gardent et al., 2017) and E2E (Novikova et al., 2017) pair meaning representations with text, much as this project pairs AL lines with English. Wiseman et al. (2017) showed that neural systems generating basketball summaries from box scores produce fluent text with many factual errors, and Puduppully et al. (2019) improved them with an explicit content plan, an intermediate structure between data and text. Pretrained text-to-text models now perform well on these benchmarks when fine-tuned on linearized inputs (Kale and Rastogi, 2020). Sports-news generation (Kanerva et al., 2019) is the closest application area: structured game events must be turned into readable commentary.

Game commentary. The nearest prior work on games is chess commentary. Jhamtani et al. (2018) collected move-by-move commentary from online forums and trained generators conditioned on board features, and Zang et al. (2019) added a chess engine to supply the analysis that a language model cannot compute. Their finding parallels mine: commentary improves when domain knowledge is supplied explicitly instead of expected from the generator. Computational work on Bridge has focused on playing rather than explaining, from the GIB program (Ginsberg, 2001) to neural bidding systems (Yeh and Lin, 2016; Rong et al., 2019). I am not aware of prior work that generates natural-language explanations of Bridge records.

Meaning representations. Semantic representations such as AMR (Banarescu et al., 2013) make the predicate–argument structure of a sentence explicit so that downstream systems need not recover it. The AL plays the same role for Bridge notation, but it also adds information that is absent from the surface, namely the conventional meaning of a call, which makes it closer to a knowledge-enriched content plan than to a pure parse (Jurafsky and Martin, 2025).

Split	Boards	Calls	Refs	Used for
Synthetic train	60	6.4	yes	writing rules
Synthetic dev	15	6.4	yes	debugging
Synthetic test	15	6.4	yes	secondary test
Real dev	40	10.5	no	tuning rules
Real test	40	12.5	yes	main test

Table 1: Dataset splits. Calls is the mean number of calls per auction (pooled over the synthetic splits). Only the real data include hands.

Evaluation. BLEU (Papineni et al., 2002) and chrF (Popović, 2015) measure overlap with a reference, and I report them using sacrebleu (Post, 2018). Overlap metrics correlate poorly with human judgments of generated text (Belz and Reiter, 2006), and they do not detect hallucination, a central failure of neural generators (Ji et al., 2023). Work on semantic accuracy checks whether each input fact is expressed in the output (Dušek and Kasner, 2020). My fact and meaning metrics follow this idea with domain-specific matching rules. I test differences with the paired bootstrap (Koehn, 2004).

3 Task Definition

The input is a Bridge board: dealer, the four hands, the auction, the final contract, the opening lead, the first trick and the number of tricks declarer won. The output is a one-paragraph English explanation for a beginner that states what happened and explains what each call means. Figure 1 shows a board from the real test set, its AL translation and the reference explanation.

The task is narrower than full commentary. Systems are not asked to judge whether a bid or play was good, to analyze the play beyond the first trick, or to take vulnerability and scoring into account.

4 Data

I use two sources: a small synthetic set written for an earlier version of the project, and real boards from Bridge Base Online (BBO) vugraph broadcasts. Table 1 summarizes the splits.

```
Raw record
  Dealer North. Hands (S.H.D.C): North .J87.QJ987.KQ532 East Q982.964.KT2.J98
                          South J654.AKQ532.654. West AKT73.T.A3.AT764
  Auction: North Pass - East Pass - South 1H - West 2H! - North 3H - East 4S -
      South 5H
         - West Double - North Pass - East Pass - South Pass
  Contract: 5HX by South Opening lead: West SA Tricks taken by declarer: 9

Artificial language (hand and play lines abbreviated)
  HAND(South, hcp=10, shape=4-6-3-0) ...
  PASS(North) PASS(East)
  OPEN(South, 1H, suit_opening, opening_strength, seat=3)
  OVERCALL(West, 2H, artificial)
  RESPOND(North, 3H, raise)
  ADVANCE(East, 4S, new_suit)
  REBID(South, 5H, agreed_suit)
  DOUBLE(West, penalty)
  PASS(North) PASS(East) PASS(South)
  CONTRACT(5H, declarer=South, doubled=X)
  LEAD(West, SA) ... WIN_TRICK(North, trick_number=1, card=H7) RESULT(South,
      contract=5H, tricks=9)

Reference (excerpt)
  North and East pass, and South opens 1H in third seat with only 10 high-card
      points but six hearts;
  third-seat openings can be light. West bids 2H in the opponents' suit, a Michaels
       cue-bid showing a
  two-suited hand with five spades and a five-card minor. North raises to 3H with
      three-card support. ...

Annotated meanings
  South 1H: third seat / light West 2H: Michaels / two-suited North 3H: raise /
      support
  West Double: penalty
```

Figure 1: A board from the real test set (real_test001). The exclamation mark marks an alerted (artificial) call. Suit symbols are written as letters here.

4.1 Synthetic Data

The synthetic set has 90 hand-written records (60 train, 15 development, 15 test) covering seven auction patterns: no-trump and suit openings, strong 2♣ openings, game contracts, doubles, redoubles, opening leads, simple play and finesses. Each record has a raw record, a gold AL translation, a reference explanation and a fact dictionary. When the parser was extended for this version, the gold AL and references were updated to the new labels, and an automatic legality check found and repaired illegal auctions and out-of-order tricks in the original set. The references follow the template wording, so this set favors the template system and is used only as a secondary test.

4.2 Real BBO Data

BBO broadcasts top-level matches as vugraph files in the LIN format, which record the dealer, all four hands, the auction with alerts, every card played and the official result. I downloaded 18 files from national and international team events; 15 contained usable boards. I wrote readers for LIN and for the PBN format. Each parsed board was replayed through the legality checker, and I kept it only if its contract, declarer and trick count agreed with the official result line in the file. Of 397 parsed boards, 392 passed and 5 were rejected. Player names, team names and broadcast commentary are discarded.

Boards were split by deal, so the two tables that played the same cards always fall in the same split, using a fixed random seed. The first 40 boards form the real development set. I used it to find and fix errors in the parser rules (Section 6.2), so it has no references.

Unseen-pattern test set. To test generalization rather than memorization of rule-covered cases, I defined an auction pattern as the sequence of non-pass calls with the non-opening side in parentheses, e.g. `1H-(2H)-3H-(4S)-5H-(X)`. The test set consists only of boards whose pattern appears nei-

ther in the synthetic training set nor in the real development set. From the remaining deals I chose one board per deal, 40 in total, greedily maximizing coverage of auction features. The resulting set has 39 distinct patterns. 30 boards are contested, 26 contain a double, 23 contain at least one alerted call (37 alerted calls in total), 13 open in third or fourth seat, 11 end in a doubled contract, 7 are slams, and in 26 the contract fails. Auctions average 12.5 calls, twice the synthetic average.

References and meaning annotations. Each test board has a reference explanation (mean 84 words) and annotated meaning facts: 105 in total, each a call such as `West 2H` with a list of acceptable keywords such as *Michaels* or *two-suited*. The references and annotations were drafted by an AI assistant (Claude) from the raw record alone, before any system was run on the test set, and I then reviewed and approved all 40. They describe what the call most likely means in expert practice, not what the parser rules say, so they can and do disagree with the AL. They have not been checked by a second Bridge player.

Facts. For every board, a fact dictionary lists the opener and opening bid, the first overcaller and overcall, the full sequence of calls, the contract, whether it was doubled, the declarer, the opening leader and lead, and declarer's trick count (9.8 facts per board on average). Facts are computed from the record, not annotated by hand.

5 Methodology

The system is a pipeline: raw record → parser → artificial language → generator. The parser and AL are the components I implemented; the generators are either my templates or an off-the-shelf language model.

5.1 Parser and Legality Checker

The parser reads players, calls (including the alert marker), contracts with doubled or redoubled status, opening leads, tricks and results, and computes each hand's high-card points and shape. A separate checker validates every record: turn order starting from the dealer, increasing bids, legal doubles and redoubles, the end of the auction, the declarer (the first player of the winning side to name the strain), the opening leader, following suit, and the winner of each trick. It runs on every record in the data as a unit test.

5.2 The Artificial Language

Each event becomes one predicate. The parser tracks the auction state (the opener, which players have acted, the last bid and whether it was artificial, which players have made takeout doubles) and uses it to choose a predicate and a meaning tag:

- **Roles.** The first bid is an `OPEN`. Later bids are a `REBID` if the player has already acted, a `RESPOND` if the player is the opener's partner, an `ADVANCE` if the player's partner has acted for the other side, and otherwise an `OVERCALL`.
- **Openings.** Openings are tagged as balanced 15–17 or 20–21 no-trump, strong artificial 2♣, weak two, preempt or suit opening, with `seat=3` or `seat=4` after passes. Alerted openings are tagged artificial.
- **Other bids.** Bids are tagged as Stayman, transfer, waiting response, raise, jump or simple overcall, no-trump overcall, cue-bid, unusual 2NT, new suit, own suit, agreed suit or second suit, or artificial when alerted.
- **Doubles.** A double of an artificial bid is lead-directing; a double of a no-trump bid is penalty; a double by responder of an overcall at the one to three level is negative; a double after partner's takeout double and an opponent's raise is responsive; a double after the doubler's partner has bid is penalty; other low-level doubles of suit bids are takeout; everything else is penalty. A redouble after a takeout double of partner's bid shows strength; otherwise it is to play.

The rules encode a simplified Standard American system. They know nothing of the agreements of a specific pair, so alerted calls outside the rule set can only be labeled `artificial`.

5.3 Generation Systems

All systems receive the same board. I compare:

1. **Generic baseline:** a fixed sentence about the deal.
2. **Literal baseline:** one sentence per call ("North bids 1NT."), then the contract and lead.
3. **Raw-copy baseline:** the raw fields copied into a sentence.
4. **Template (beginner, expert):** hand-written templates over the AL, in a beginner style that explains each tag and a shorter expert style.

5. **LLM, raw:** a language model given the raw record, including hands, play and result.
6. **LLM, AL:** the same model given the AL lines.

The two LLM conditions differ only in the user message. The system prompt asks the model to act as a Bridge teacher, go through the auction in order explaining each call, state the contract, declarer, opening lead, first trick and result, use only information in the record, and write one paragraph of prose. I use two Qwen3 models (Yang et al., 2025) run locally with Ollama in its default 4-bit quantization: Qwen3-4B-Instruct-2507, which does not produce reasoning traces, and Qwen3-1.7B with thinking disabled. A third candidate, the Qwen3-4B build available locally, could not be run without reasoning and was not used. Decoding is greedy (temperature 0, seed 0) with up to 700 output tokens. Every output is cached with a hash of its prompt, so all results can be reproduced without rerunning the models.

6 Experiments

6.1 Metrics

Fact accuracy. The fraction of facts the output states in the right role. A value counts only if it appears in a sentence that also contains a cue for its role (e.g. *open* or *start* for the opener, *lead* or *led* for the opening lead, *contract* or *declarer* for the declarer), or the value of its paired fact ("South bids 1♡" states both opener and opening bid). Bids and cards are matched in several spellings (`3NT`, *three no-trump*, *six of hearts*). The call-sequence fact is correct only if every non-pass call is mentioned in order.

Meaning recall. The fraction of annotated meanings explained: the calling player and one of the keywords must appear in the same sentence. On the synthetic set, meaning facts are read from the gold AL (e.g. *balanced* and *15-17* for a 1NT opening), which favors systems that copy AL tags.

Unsupported bids. The number of bids per output that the auction never contains, counted only after a calling word such as *bids*, *opens* or *raises to*, so that cards and hand shapes ("two hearts") are not counted. This is a hallucination check; lower is better.

Readability and overlap. Flesch–Kincaid grade level (Kincaid et al., 1975) (lower is easier), output length, and corpus BLEU and chrF against the reference, computed with sacrebleu. On the synthetic set I also report AL exact match against the gold AL.

Significance. For the main comparison I compute per-board differences between the AL and raw conditions of the same model and a 95% confidence interval from 10,000 paired bootstrap resamples over boards (Koehn, 2004).

6.2 Development Process

The system has no trained parameters: the rules and templates were written from the synthetic training set and tuned on the development sets. Because nothing is fitted to data, I did not use cross-validation. Reviewing template output on the real development set exposed six rule errors, all fixed before the test set was built: a double was not treated as a side entering the auction; no-trump bids were classified as cue-bids; natural no-trump overcalls were described as suits; cue-bids by responder or advancer were called new suits; rebidding an agreed suit was described wrongly; and responsive doubles were classified as penalty. No rule was changed after looking at the test set.

One change to the scorer was made after seeing test outputs. The first version of fact accuracy required a role word in the sentence, and it marked correct sentences such as "East started the auction by bidding one club" and "East led with the two of clubs" as wrong. I added the paired-fact rule and the cues *start*, *first* and *led*, and removed the generic cue *bids* for overcalls. The change applies equally to all systems, and Section 7 reports the main comparison under both versions.

7 Results

Table 2 gives the main results and Table 3 the paired comparisons.

The AL improves meaning without costing facts. With the same model and instructions, AL input raises meaning recall from 0.21 to 0.63 for the 4B model and from 0.05 to 0.31 for the 1.7B model. Both intervals exclude zero. On a per-board basis the 4B model explains more meanings with AL input on 28 boards and fewer on none; the 1.7B model is better on 21 and worse on 2. Fact accuracy does not change measurably for either model. Under the original, stricter fact scorer, AL input helped facts as well (4B: 0.63 raw, 0.85 AL;

System	Facts ↑	Meaning ↑	Unsup. ↓	FK grade	Words	BLEU	chrF
Generic baseline	0.00	0.00	0.00	4.8	10	0.0	7.2
Literal baseline	0.86	0.03	0.00	0.5	61	12.7	36.3
Raw-copy baseline	0.88	0.04	0.00	12.7	62	1.4	28.0
Qwen3-1.7B, raw	0.82	0.05	0.07	5.1	113	7.3	41.8
Qwen3-1.7B, AL	0.80	0.31	0.00	5.7	122	7.7	42.4
Qwen3-4B, raw	0.90	0.21	0.00	7.5	203	4.2	40.5
Qwen3-4B, AL	0.90	**0.63**	0.00	8.1	233	5.6	41.7
Template (expert), AL	**1.00**	0.45	0.00	1.4	117	**18.4**	51.2
Template (beginner), AL	**1.00**	**0.64**	0.00	3.5	206	16.8	**52.1**

Table 2: Results on the real test set (40 boards, all auction patterns unseen). Facts and Meaning are mean fractions per board; Unsup. is unsupported bids per output; BLEU and chrF are corpus-level. References average 84 words.

Model	Metric	AL − raw	95% CI
4B	Meaning	$+0.42$	$[+0.31, +0.53]$
4B	Facts	$+0.00$	$[-0.04, +0.05]$
4B	chrF	$+1.2$	$[+0.3, +2.1]$
1.7B	Meaning	$+0.26$	$[+0.16, +0.36]$
1.7B	Facts	-0.02	$[-0.08, +0.05]$
1.7B	chrF	$+0.5$	$[-0.4, +1.3]$

Table 3: Paired differences between AL and raw input on the real test set, with paired bootstrap confidence intervals (chrF is sentence-level here).

System	Facts	Meaning	chrF
Literal baseline	0.97	0.00	46.2
Raw-copy baseline	0.99	0.00	34.4
Qwen3-1.7B, raw	0.95	0.08	45.0
Qwen3-1.7B, AL	0.88	0.32	50.0
Qwen3-4B, raw	0.97	0.17	48.4
Qwen3-4B, AL	0.94	0.83	46.6
Template (beginner), AL	1.00	1.00	99.6

Table 4: Results on the synthetic test set (15 examples). Template scores are inflated because the references follow the template wording.

1.7B: 0.73 and 0.75; beginner template 1.00); the relaxed scorer mainly credits correct paraphrases in raw-input outputs, so the change works against my hypothesis, not for it.

Hallucinated bids are rare. Only the 1.7B model on raw input invents bids, at 0.07 per output (one invented bid in each of three outputs). The larger problem with raw input is not invented calls but wrong explanations of real ones (Section 8).

The template system matches the best model. The beginner template states every fact and explains 64% of meanings. The difference from the 4B model with AL input is -0.01 (95% CI $[-0.08, +0.05]$), so neither is better on meaning. The template is far easier to read (grade 3.5 against 8.1) and closer to the references (chrF 52.1 against 41.7). The LLMs write long paragraphs, 2.5 times the reference length, which lowers BLEU for all of them.

Baselines. The literal and raw-copy baselines state most facts (0.86 and 0.88) but explain almost nothing, which confirms that fact accuracy alone does not measure the task. The generic baseline scores zero on both.

Synthetic test set. Table 4 shows the same pattern on the 15 synthetic test examples: AL input raises 4B meaning recall from 0.17 to 0.83 and 1.7B from 0.08 to 0.32. Template scores are near-perfect because the references follow the template wording and meaning facts come from the gold AL. The parser's AL matches the gold AL exactly on all 15 examples, which shows consistency on the seven synthetic patterns rather than general parsing accuracy.

8 Discussion

8.1 What the AL Contributes

Given the raw record, the models read bids literally. On the board in Figure 1, the 4B model with raw input says that South's 1♡ showed "at least one heart" and that West's 2♡ "meant they had at least two hearts", reading the bid level as a card count. West's bid is a Michaels cue-bid in the opponents' suit and says nothing about hearts. Given the AL line `OVERCALL(West, 2H, artificial)`, the same model writes that West made "an artificial overcall of two hearts, not showing strength in hearts but signaling a different kind of hand". That is still not the full meaning, but it is correct. The template writes that the bid is "an alerted artificial bid whose meaning depends on the partnership's system". The AL gives the model the facts a begin-

ner's teacher would start from, partnership role and the type of each call, and the model passes them on.

8.2 The Model Adds Little Knowledge

Comparing the 4B model with AL input against the beginner template on the 105 meaning facts, both explain 62, the model alone explains 4, the template alone explains 6, and neither explains 33. The model almost never recovers a meaning that is not already in the AL. Its gain over raw input therefore comes from the representation, and its ceiling is the representation's coverage. This is the same lesson as in chess commentary (Zang et al., 2019): a small language model verbalizes knowledge well but does not supply it.

8.3 Error Analysis

Missing conventions. The beginner template misses 37 of the 105 meanings. Most of them (29) involve conventions or treatments that the simplified Standard American rules do not model: Multi 2♢ openings and the pass-or-correct responses to them (6 facts), weak no-trump openings (3), two-suited cue-bids such as Michaels (3), and single cases of Lebensohl, key-card asking, inverted minor raises, reverses and strong club openings. For alerted calls the AL says only `artificial`, which is honest but uninformative. Five misses are doubles where the rule-based classification disagrees with the expert reading. Three takeout doubles are labeled lead-directing (two) or penalty (one); a penalty double is labeled lead-directing; and a lead-directing double is labeled penalty. The rule that a double of an artificial bid is lead-directing fails most often: over an opponent's Multi 2♢, for example, a double is normally takeout.

Metric artifacts. Three misses are seat notes that the template does state, but in a separate sentence without the player's name ("This opening comes in third seat…"), which the same-sentence rule does not credit. The fact scorer has the opposite weakness: the call-sequence fact checks that calls appear in order but not who made them, and I observed outputs, especially from the 1.7B model, that credit a call to the wrong player without penalty.

Readability. The LLMs write at grade 7.5–8.1 against 3.5 for the beginner template, and produce 2.5 times the reference length. The instruction to explain every call encourages long, clause-heavy sentences. A beginner reader would likely prefer the template, but only a human study can show this.

Smaller model. The 1.7B model writes about half as much as the 4B model and, in longer auctions, skips calls and confuses players even with AL input. On one 15-call board it says North opened 1NT and West doubled, when West opened and North doubled, and it omits the redouble and East's 2♣. This explains its low meaning recall in both conditions.

8.4 Comparison with Previous Work and the State of the Art

There is no previous Bridge commentary system to compare against directly, so the baselines and the template system serve as reference points. Relative to data-to-text work, the findings match two established results: fluent neural output is not faithful by default (Wiseman et al., 2017), and an explicit intermediate plan improves content selection (Puduppully et al., 2019). The present system is far from a state-of-the-art commentator. Approaching one would need (1) a much larger aligned corpus, which could be built from BBO's archive of millions of boards with model-drafted, expert-reviewed references; (2) knowledge of each pair's system, which BBO records partially as convention cards and alert explanations; (3) larger or fine-tuned generators (Kale and Rastogi, 2020) trained on AL–English pairs; (4) a double-dummy solver, as GIB uses (Ginsberg, 2001), to judge whether bids and plays were good; and (5) human evaluation by Bridge players and beginners.

9 Conclusion

I asked whether translating Bridge records into a controlled artificial language helps a language model explain them. On 40 real boards with unseen auction patterns, it does: with the AL, a 4B model explains three times as many bid meanings, without losing facts and without inventing bids. A template generator reading the same AL performs as well as the best model on meaning, states every fact and is easier to read. The model's explanations are bounded by what the AL encodes, so improving the representation, above all its coverage of conventions, is the most direct way to improve the output.

10 Future Work

The most valuable next step is extending the AL with partnership agreements, for example reading alert explanations and convention cards from BBO records so that `artificial` can become `Multi` or `Michaels`. A human evaluation should replace the automatic proxies: Bridge players scoring correctness and beginners scoring readability and whether they could reconstruct the auction from the text. The fact scorer should check who made each call. A formal grammar for the AL would let malformed representations be detected automatically, and a reverse task, rebuilding the record from the English, would measure reconstructability directly. Finally, larger models and fine-tuning on AL–English pairs would show whether the AL's benefit shrinks as models know more Bridge.

Limitations

The real test set has 40 boards, so confidence intervals are wide for all but the largest effects. The references and meaning annotations were drafted by an AI assistant and reviewed by one author, not by independent Bridge experts, and meaning recall depends on their keyword lists. All metrics are automatic and based on word matching; they can miss correct paraphrases and credit sentences that mention the right words in a wrong claim. One scorer change was made after viewing test outputs (Section 6.2). Only two small quantized models were tested, with one prompt, and larger models may know more conventions. The bidding rules assume a simplified Standard American system and ignore vulnerability and scoring.

Ethical Considerations

The real boards come from public BBO vugraph broadcasts. Player and team names and broadcast commentary are discarded, and only the cards, calls and results are kept. The models run locally, so no data is sent to external services. The main risk of an explanation system is false confidence: a fluent explanation of a convention the pair did not play would mislead a beginner. The AL therefore labels unknown alerted calls as artificial rather than guessing, and unsupported bids are counted as errors.

Contributions and AI Assistance

This is a single-author project. I defined the task and research questions, chose the data sources and models, and reviewed and approved all test references. An AI coding assistant (Claude, Anthropic) was used to help write the code, draft the test references and annotations, and draft parts of this paper; all of its output was reviewed by me.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133. Association for Computational Linguistics.

Matthew L. Ginsberg. 2001. GIB: Imperfect information in a computationally challenging game. *Journal of Artificial Intelligence Research*, 14:303–358.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1661–1671.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Daniel Jurafsky and James H. Martin. 2025. Speech and language processing. https://web.stanford.edu/~jurafsky/slp3/.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102.

Jenna Kanerva, Samuel Rönnqvist, Riina Kekki, Tapio Salakoski, and Filip Ginter. 2019. Template-free data-to-text generation of Finnish sports news. In

Proceedings of the 22nd Nordic Conference on Computational Linguistics, pages 242–252.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report Research Branch Report 8-75, Naval Technical Training Command.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6908–6915.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Jiang Rong, Tao Qin, and Bo An. 2019. Competitive bridge bidding with deep neural networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 16–24.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263. Association for Computational Linguistics.

An Yang and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chih-Kuan Yeh and Hsuan-Tien Lin. 2016. Automatic bridge bidding using deep reinforcement learning. In *Proceedings of the 22nd European Conference on Artificial Intelligence*, pages 1362–1369.

Hongyu Zang, Zhiwei Yu, and Xiaojun Wan. 2019. Automated chess commentator powered by neural chess engine. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5952–5961.

A Prompts

System prompt, shared by all LLM conditions:

```
You are a bridge teacher. Explain the
    bridge deal record you are given to
     a beginner in plain English. Go
    through the auction in order and
    say what each bid, double or
    redouble means. Then state the
    final contract and who is declarer,
     the opening lead, what happens on
    the first trick, and how many
    tricks declarer took, if the record
     says. Use only the information in
    the record and do not invent bids,
    cards or results. Write one
    paragraph of plain prose with no
    headings, lists or markdown.
```

The raw condition's user message starts with "Bridge record:" followed by the dealer, hands, auction, contract, opening lead, play and trick count, one field per line. The AL condition's user message starts with "Bridge record, one event per line in a structured meaning representation:" followed by the AL lines.

B Synthetic Example

```
Raw:
Auction: North 1NT - East Pass - South
     3NT - West Pass - North Pass -
    East Pass
Contract: 3NT by North
Opening lead: West S4

Artificial language:
OPEN(North, 1NT, balanced, 15-17_HCP)
PASS(East)
RAISE_TO_GAME(South, 3NT)
PASS(West)
PASS(North)
PASS(East)
CONTRACT(3NT, declarer=North)
LEAD(West, S4)

English (beginner template):
North opens 1NT, showing a balanced
    hand with about 15--17 high-card
    points. East passes. South raises
    directly to 3NT, which is a game-
    level no-trump contract. West
    passes. North passes. East passes.
    The final contract is 3NT by North.
     West leads the four of spades to
    begin the play.
```