SemEval11 - Dataset Analysis Report

Prepared: Generated from

Executive summary

This report organizes the quantitative and lexical analyses performed on the SemEval11 syllogism dataset (960 samples). It highlights dataset size, distributions of logical validity and real-world plausibility, relationships between those labels, text-length statistics, vocabulary and quantifier usage, and entity frequency broken down by the four validity × plausibility quadrants.

---

1. Dataset overview

- **Total number of syllogisms: 960**

1.1 Distribution by logical validity

- **Valid:** 480 (50.0%)
- **Invalid:** 480 (50.0%)

1.2 Distribution by real-world plausibility

- **Plausible:** 474 (49.4%)
- **Implausible:** 486 (50.6%)

---

2. Validity vs. Plausibility (contingency)

Counts:

Validity	Implausible	Plausible	Total
Invalid	246	234	480
Valid	240	240	480
Total	**486**	**474**	**960**

Percentages of total dataset:

Validity	Implausible	Plausible
Invalid	25.6%	24.4%
Valid	25.0%	25.0%

> Observation: the dataset is balanced across validity (50/50) and near-balanced across plausibility; each quadrant contains roughly ~240 examples (range 234–246).

---

3. Text length analysis

Overall length statistics

Metric	Char_count	Word_count
count	960	960
mean	134.54	25.17

std	25.04	4.55
min	70.00	14.00
25%	117.00	22.00
50% (median)	132.00	25.00
75%	150.25	28.00
max	234.00	40.00

Average word count grouped by category

Validity	plausibility=False	plausibility=True
Invalid (False)	24.19	24.36
Valid (True)	25.81	26.32

> Note: Valid syllogisms tend to be slightly longer (in words) than invalid ones; plausible examples are also marginally longer.

---

4. Vocabulary (Top words)

Top 25 most common (non-stop) words and counts:

- every: 525
- single: 389
- animals: 187

- also: 177
- anything: 176
- case: 176
- fish: 173
- therefore: 171
- mammals: 167
- animal: 163
- follows: 144
- mammal: 143
- true: 125
- things: 122
- birds: 112
- dogs: 109
- dog: 104
- consequently: 97
- human: 88
- cats: 84
- bird: 83

- cars: 78
- everything: 76
- portion: 75
- people: 75

> These high-frequency lexical items indicate heavy use of universal quantifiers (`every`, `all`/`every`-type language) and recurring domain concepts like animals, mammals, birds, and people.

---

5. Quantifier frequency

Quantifier	Frequency
some	679
no	567
every	525
all	258
a portion of	73
any	60
not a single	56
a few	55
nothing that is	51

at least one	37
a number of	28
each	13

> Takeaway: `some`, `no`, and `every` dominate; both universal and existential quantifiers are well represented which is important for logical/semantic modeling.

---

6. Entity (noun) frequency per quadrant

Top 10 entities for each quadrant (counts are raw occurrences within that quadrant):

Valid & Plausible (240 samples)

Rank	Entity	Count
1	animals	47
2	mammals	31
3	animal	29
4	birds	21
5	person	19
6	bird	19
7	cars	18
8	dogs	18

9	trees	17
10	people	17

Valid & Implausible (240 samples)

Rank	Entity	Count
1	animals	44
2	water	26
3	animal	24
4	mammals	24
5	planet	22
6	plants	20
7	planets	20
8	dogs	19
9	fish	19
10	humans	18

Invalid & Plausible (234 samples)

Rank	Entity	Count
1	mammals	61
2	animal	51

3	animals	49
4	dog	49
5	dogs	44
6	cats	41
7	birds	39
8	people	32
9	fish	29
10	mammal	28

Invalid & Implausible (246 samples)

Rank	Entity	Count
1	animal	54
2	mammals	51
3	animals	47
4	birds	35
5	dog	32
6	fish	31
7	cat	30
8	bird	29

9	dogs	28
10	cars	27

Observations:

- Animals-related entities (animals, animal, mammals, mammal, birds, fish, dogs) are pervasive across all quadrants.
- Some domain shifts appear across quadrants (e.g., `water`, `planet(s)` appear in Valid & Implausible), which could indicate particular semantic templates used to create implausible but logically valid examples.

---

7. Suggested next steps / uses

1. **Modeling experiments:** Use quadrant-balanced cross-validation splits to ensure models see a representative mix of validity and plausibility.
2. **Feature engineering:** Quantifier counts and presence (some/no/every) could be strong features; consider dependency parse patterns for quantifier scope.
3. **Data augmentation:** If you need more examples for under-represented entity/quantifier combinations, targeted augmentation can help.
4. **Error analysis:** Inspect cases where models confuse validity vs. plausibility - sample from each quadrant for manual analysis.