

The Connective *but* in Informal Mathematical Proofs: A Corpus-Based Study

Rinta Yamada^{1,2}, Sakiko Yamasaki², Risako Ando², Takayuki Amamoto²,
Ryota Akiyoshi^{1,2}, Koji Mineshima²

¹The University of Electro-Communications, Chofu, Tokyo, Japan

²Keio University, Minato, Tokyo, Japan

{rinta.yamada.cs, sakiko.yamasaki8008, georg.logic}@gmail.com

{risakochaan, amamoto, minesima}@keio.jp

Abstract

This study presents a focused case study of the explicit connective *but* in informal mathematical proofs. We propose an annotation scheme and analyze 261 occurrences found in 225 proofs from textbooks in logic, set theory, arithmetic, Boolean algebra, and formal semantics. The P-Label distinguishes proof-relevant content—material that forms an essential part of the proof—from metatextual comments on the proof, with contradiction-completing cases marked separately within the proof-relevant class. An optional R-Label records contrast or exception. Of the 261 occurrences, 221 (84.7%) were classified as proof-relevant. Three illustrative case studies with corresponding Lean formalizations show that informal transitions may align with formal operations but need not correspond to single proof lines, clarifying why proof relevance is not defined by line-level correspondence. In grouped five-fold prediction experiments, GPT-5.6 Terra attains 58.6% hierarchical accuracy from the target sentence. In these fixed runs, full-proof input yielded lower hierarchical accuracy in all four comparisons, with observed decreases of 15.3–26.1 points for Qwen3-8B. The study characterizes one connective rather than the full inventory of discourse relations in mathematics, providing a basis for future comparisons with other explicit connectives and implicit relations in informal mathematical proofs.

1 Introduction

Informal proofs in mathematical textbooks and research papers are written in a hybrid language that combines natural language with mathematical notation, exhibiting domain-specific vocabulary and syntactic structures (Ranta, 1995; Ganesalingam, 2013). Beyond the level of sentences, recognizing the relationships between them — namely, the *discourse relations* in mathematical reasoning — is crucial for understanding how arguments are structured and how theorems are justified. However,

discourse relations specific to mathematical proofs have received relatively little attention in computational linguistics or in other related fields.

Mathematical discourse employs logical connectives such as *and*, *or*, and *if*, inferential markers such as *therefore*, *since*, and *because*, and other expressions that help readers follow the organization of a proof. Among the latter, *but* is a useful case for examining how an explicit lexical marker interacts with mathematical reasoning. While concession has been studied extensively in theoretical linguistics and discourse analysis (Zufferey and Degand, 2024), corpus-based research on *but* in proof discourse remains limited. Understanding this structure is not only of linguistic interest but also crucial for advancing research on automated proof formalization (Szegedy, 2020; Weng et al., 2025) and informalization (Kaliszyk et al., 2017; Asperti et al., 2026), where the goal is to reconstruct formal proofs from informal mathematical texts and vice versa.

In this study, we propose an annotation scheme and a corpus-based analysis of mathematical textbooks, focusing on the connective *but* in informal mathematical proofs. The study is deliberately framed as a case study of this single word form. We analyze explicit tokens of *but/But*; we do not annotate relations expressed by other connectives, compare connective frequencies, or recover implicit discourse relations without an overt marker. Through this focused analysis, we aim to characterize the functions of *but* in mathematical reasoning and establish a basis for comparative studies.

The main findings of this study can be summarized as follows.

First, we construct and annotate a focused corpus of 261 occurrences of *but* in 225 proofs from six mathematical textbooks. We introduce a hierarchical scheme that separates the proof-level contribution of the associated material from optional contrastive and exceptive uses (Sections 3 and 4.1).

Second, distributional and qualitative analyses show that 221 occurrences are proof-relevant and that 158 of these carry no R-Label, while individual cases illustrate diverse proof-advancing functions of *but* (Sections 4.1 and 4.2).

Third, three illustrative case studies with corresponding formalizations in the proof assistant Lean¹ illustrate two relatively direct alignments and one mismatch, clarifying why line-level alignment does not define proof relevance (Section 4.3).

Finally, grouped five-fold experiments with a linear baseline, GPT-5.6 Terra, and the open-weight Qwen3-8B model provide initial benchmarks for predicting the hierarchical labels and test whether a full proof is more useful than the target sentence alone (Section 5).

2 Discourse Relations and Informal Mathematical Proofs

2.1 Discourse Relations in Natural Language

Discourse relations have been modeled and annotated in frameworks such as Rhetorical Structure Theory (RST) (Carlson et al., 2002), Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), and the Penn Discourse Treebank (PDTB) (Prasad et al., 2014; Webber et al., 2019). While RST and SDRT model rhetorical or discourse-semantic structure, PDTB takes an argument-based approach: it identifies explicit or implicit relations between two discourse units, *Arg1* and *Arg2*, and assigns a relation label.

While the PDTB framework characterizes how discourse units are related, informal mathematical proof raises a complementary proof-level question: whether the material associated with a connective supplies a step required to complete the proof. A local relation alone does not record this contribution to the derivation as a whole. Example (1), adapted from Kocurek and Pavese (2022, p. 415), illustrates this distinction.

- (1) Suppose p is the largest prime, and let q be a prime divisor of $p! + 1$. Then $q > p$, since no prime at most p divides $p! + 1$. **But** this contradicts the assumption that p is the largest prime. Therefore, there is no largest prime number.

Here the clause introduced by *But* completes the contradiction needed to reject the initial assumption, a proof-level contribution not recorded by

a local relation alone. We distinguish an optional relational-use subtype associated with *but* (R-Label) from the proof-level contribution of its associated material (P-Label). Section 3 defines these two annotation dimensions.

2.2 Mathematical Discourse and Formalization

Formal and computational studies of mathematical language have drawn on formal grammar and formal semantics such as Discourse Representation Theory to analyze linguistic and discourse phenomena in mathematical texts (Ranta, 1995; Ganesalingam, 2013; Matsuzaki et al., 2017). The Naproche project similarly uses a controlled natural language and logical representation to support the verification of natural-language proofs (Cramer et al., 2009). Together, these approaches demonstrate the importance of linguistic structure for recovering mathematical content from text.

Their use of discourse representations, however, should be distinguished from annotating the rhetorical organization of an informal proof. These approaches are not primarily designed to record how an overt connective guides a reader through the progression of a proof, or whether its associated material belongs to the derivation rather than to the surrounding exposition. This organizational contribution may disappear in formal representation because a connective such as *but* need not have a direct formal counterpart.

The present study addresses this intermediate layer between discourse-relation annotation and formal proof representation. It distinguishes the proof-level contribution of the material associated with *but* from any local contrastive or exceptive meaning. This perspective complements formalization: the connective may signal an inferential transition even when it does not correspond to a single line or expression in a formal proof.

3 Annotation Scheme for *but*

3.1 Label Definition and Examples

This section introduces a hierarchical annotation scheme for occurrences of the connective *but* in proofs. At the primary level, a Proof-Relevance Label (P-Label) classifies the role of the material associated with *but* in the proof. At the optional second level, a Relation Label (R-Label) records a relational-use subtype when one is identified. Every instance therefore receives a P-Label, but re-

¹<https://lean-lang.org/>

ceives an R-Label only when it expresses CONTRAST or EXCEPTION.

3.1.1 Primary Proof-Relevance Labels

At the primary level, we define *Proof-Relevance Labels* (*P-Labels*, for short): PROOF, PROOF-NEG, and META. The P-Label concerns the content associated with *but*, not the word itself: PROOF and PROOF-NEG jointly mark proof-advancing transitions, whereas META marks content that does not advance the derivation. Beyond distinguishing contradiction-completing transitions as PROOF-NEG, we do not annotate finer transition types.

The label PROOF is assigned when the content associated with *but* forms part of the reasoning used to establish the conclusion of the proof.

(2) **Label:** PROOF

[...] we can conclude that $a \in B$. **But** $a \notin B \setminus C$, so it follows that $a \in C$. (Velleman, 2019, p. 111)

The label PROOF-NEG is a more specific proof-relevant category. It is assigned when the associated content states or completes a contradiction that is used to reject an assumption or derive a negated conclusion.

(3) **Label:** PROOF-NEG, CONTRAST

Suppose $x^2 + y = 13$ and $y \neq 4$. Suppose $x = 3$. [...] **But** this contradicts the fact that $y \neq 4$. Therefore $x \neq 3$. (Velleman, 2019, p. 104)

These proofs would become logically incomplete if the content of the clauses initiated by *But* were removed. Note that this deletion test concerns the clause content, not deletion of the connective alone. In particular, the latter clause states the contradiction used to derive the negated conclusion. Thus, (2) is marked with the label PROOF, and (3) has the label PROOF-NEG.

On the other hand, a proof may contain *but* that is used in such contexts as additional remarks, qualifications, caveats, or side comments that are not essential for its validity. We classify this use with the label META: after removing the associated comment, the remaining propositions and inference steps still suffice for the proof. In (4) below, the *but*-clause does not contribute a necessary step to the proof.

(4) **Label:** META

A function $f: \omega \rightarrow A$ for which $f(n^+) < f(n)$ (for all $n \in \omega$) is sometimes called

a descending chain, **but** this terminology should not be confused with other uses made of the word “chain.” (Enderton, 1977, Proof of Theorem 7B, p. 173)

Annotators apply these decisions in the order META, PROOF-NEG, and PROOF; the full decision procedure is given in Appendix A.

3.1.2 Optional Relational-Use Subtypes

The two representative uses of the connective *but* in English are *contrast* and *exception* (Huddleston and Pullum, 2002, p. 1310).

(5) John is tall **but** Paul is short. (Contrast)

(6) Everyone **but** him came. (Exception)

At the optional second level of the hierarchy, we use two labels for relational-use subtypes associated with *but*, namely, CONTRAST and EXCEPTION. These *Relation Labels* (*R-Labels*, for short) supplement the primary P-Label rather than forming an independent classification. The two levels encode different dimensions and are not mutually exclusive alternatives. The P-Label records the proof-level status of the material associated with *but*, whereas the R-Label optionally records whether the same occurrence also instantiates CONTRAST or EXCEPTION. Thus, an occurrence may advance the proof and simultaneously express a contrast, yielding PROOF, CONTRAST.

The term R-Label is used for convenience: the two subtypes are not homogeneous discourse relations. CONTRAST normally relates two discourse arguments, whereas EXCEPTION also includes phrasal exceptives such as (6), which need not connect two clauses and therefore fall outside the usual inventory of PDTB-style explicit discourse relations. We include denial-of-expectation and concessive readings under CONTRAST. Although Contrast and Concession are often distinguished (Kehler, 2002; Winterstein, 2012), including in PDTB 3.0 (Webber et al., 2019), their boundary is difficult to apply consistently in practice; we therefore do not annotate Concession separately. No correction uses, which are most characteristic of dialogue, were found in the present corpus of expository mathematical proofs. Descriptions such as goal shift or inferential continuation concern the proof-functional dimension represented by the P-Label; they are not alternatives to CONTRAST or EXCEPTION and may co-occur with an R-Label.

The CONTRAST label is assigned when the material surrounding *but* explicitly aligns two alternatives and highlights an opposition or difference between them. Both alternatives must be identifiable from the text.

The EXCEPTION label is assigned only when a generalization or shared frame remains in force while a particular member, case, parameter, or respect is excluded or replaced. The occurrence must permit a meaning-preserving paraphrase with *except* or *except for*, as in “the same proof except for the choice of parameter.”

(7) **Label:** PROOF, CONTRAST

If E had an element not in D , the complement of such an element would be an upper bound of D **but** not of E . (Halmos, 1974, Proof of Lemma 1, p. 61)

(8) **Label:** META, EXCEPTION

This is the same as the pigeonhole principle, **but** for an arbitrary finite set A instead of a natural number. (Enderton, 1977, Proof of Theorem 6C, p. 135)

Example (7) contrasts the upper-bound status of D and E . In Example (8), *but* marks the parameter that changes within an otherwise shared proof pattern and allows an *except for* paraphrase; we therefore assign CONTRAST and EXCEPTION, respectively.

In informal mathematical proofs, *but* is also commonly used in a way that does not meet either operational R-Label criterion. Consider the following example:

(9) **Label:** PROOF

Suppose $\neg R$. Suppose P . Since P and $P \rightarrow (Q \rightarrow R)$, it follows that $Q \rightarrow R$. **But** then, since $\neg R$, we can conclude $\neg Q$. Thus, $P \rightarrow \neg Q$. Therefore $\neg R \rightarrow (P \rightarrow \neg Q)$. (Velleman, 2019, p. 110)

This occurrence advances the deduction and therefore receives the P-Label PROOF. However, it does not meet either operational R-Label criterion: explicit opposing properties P_1 and P_2 are difficult to identify, and *but* cannot be paraphrased as *except*. Hence no R-Label is added. Proof progression is not an alternative to Contrast: an occurrence that both advances the deduction and expresses an identifiable opposition receives PROOF, CONTRAST, as in Example (7).

3.1.3 Supplementary Annotation

In addition to the labels defined above, each occurrence of *but* was further annotated for its position and form. Occurrences of capitalized *But* at the beginning of a sentence were labeled INITIAL, while lowercase *but* appearing within a sentence was labeled MEDIAL.

Compound expressions with *but* (e.g., *but then*, *but since*) were recorded separately.

3.2 Corpus Overview

The corpus is based on six textbook sources spanning several domains, including logic, set theory, arithmetic, Boolean algebra, and formal semantics. Logic, set theory, and arithmetic are represented in Velleman (2019), Partee et al. (1990), Enderton (2001), and Enderton (1977); Boolean algebra in Halmos (1974) and Partee et al. (1990); and formal semantics in Landman (1991).

To construct the corpus, we identified mathematical statements, including exercise problems, in searchable electronic versions of the sources and searched their associated proof passages for the word forms *but* and sentence-initial *But*. The search covered the full set of target pages in each source. For each match, the corresponding statement and complete proof were OCR-transcribed from the source pages and manually checked against the originals. Proof length was measured over the 225 unique proofs, rather than the 261 target occurrences, using these manually checked OCR transcriptions. We counted Unicode word-like tokens after whitespace normalization and correction of OCR line-break hyphenation; punctuation and standalone mathematical operators were excluded. The median proof length was 124 tokens (IQR 60–196; range 14–696).

Across these sources, 225 proofs containing at least one occurrence of *but* were identified, totaling 261 instances. When multiple occurrences of *but* appeared within a single proof, each was indexed according to its order of appearance and assigned an individual label. Occurrences of *but* in statements were excluded from the count, so only those appearing within proofs were annotated. The extraction targeted only the word forms *but* and sentence-initial *But*. Other connectives (e.g., *and*, *however*, and *although*) and discourse relations without an overt connective were not collected or annotated. Their exclusion reflects the case-study design and should not be interpreted as evidence

P-Label	Contrast	Exception	No R-Label	Total
PROOF	41	5	117	163
PROOF-NEG	17	0	41	58
META	11	5	24	40
Total	69	10	182	261

Table 1: Cross-distribution of optional relational-use subtypes and proof-relevance labels.

that they are absent or unimportant in mathematical proofs.

Annotation procedure. Six annotators, all trained in logic, worked in varying pairs, with two independently labeling each instance after inspecting the complete proof. During this phase, they could record an *uncertain* or *other* response for difficult cases. All disagreements and flagged cases were then reviewed by the six annotators, and every instance received an adjudicated P-Label and a decision on whether an R-Label applied. Agreement was calculated from the independent responses, whereas the distributional analyses and prediction experiments use the adjudicated labels for all 261 instances.

Inter-annotator agreement. Because annotator pairs varied across instances, we report Fleiss’ κ , based on the pooled category distribution across the six annotators. In the primary analysis, *uncertain* and *other* responses were retained as distinct annotation outcomes rather than removing difficult cases. Agreement was 210/261 (80.5%; $\kappa = .615$) for the P-Label and 176/261 (67.4%; $\kappa = .554$) for the complete hierarchical label, where a case was represented either by its P-Label alone or by the P- and R-Labels together.

R-Label reliability was also evaluated separately. Among the 257 cases with resolved R responses, agreement was 220/257 (85.6%; $\kappa = .661$) on whether to assign any R-Label and 215/257 (83.7%; $\kappa = .636$) on the exact three-way decision among CONTRAST, EXCEPTION, and no R-Label. Restricting the latter analysis to the 210 cases with matching valid P-Labels (209 after excluding one unresolved R response) yielded 176/209 agreement (84.2%; $\kappa = .643$). Appendix A gives sensitivity analyses, the R-Label confusion matrix, an audit of the 10 adjudicated EXCEPTION cases, and qualitative disagreement patterns.

4 Annotation Results and Discussion

4.1 Distribution

We report the adjudicated distribution of all 261 instances from the six textbooks.² Table 1 cross-tabulates the primary P-Labels and optional R-Labels.

As shown in Table 1, 221 occurrences (84.7%) were proof-relevant (PROOF or PROOF-NEG) and thus mark proof-advancing transitions in our coarse-grained sense; META uses accounted for 15.3%. Within the proof-relevant category, PROOF cases (163/221; 73.8%) outnumbered PROOF-NEG cases (58/221; 26.2%), showing that general PROOF cases were more common than contradiction-completing PROOF-NEG cases.

The same table shows the distribution of optional relational-use subtypes. The CONTRAST R-Label accounted for 26.4% (69 instances) and EXCEPTION for 3.8% (10 instances). Notably, 69.7% of the instances (182 in total) received no R-Label and were therefore represented by their P-Label alone. Of the 221 proof-advancing transitions, 158 (71.5%; 60.5% of all occurrences) were P-only, carrying neither a CONTRAST nor an EXCEPTION R-Label.

The No R-Label outcome does not mean that these occurrences are unclassified; it means only that neither optional relational-use subtype was identified. A finer-grained pragmatic inventory might subdivide some of these cases, but such distinctions would be orthogonal to the proof-functional classification and would not alter whether the associated content advances the proof. The recurring P-only patterns examined below therefore support the importance of proof-level discourse functions, although comparison with non-mathematical expository discourse would be needed to establish their mathematical specificity.

Table 2 reports the number and median length of the unique proofs in each source, together with the source-level label distribution. Source-level median proof lengths range from 98 tokens in Landman to 161 in Enderton’s set-theory text. Proof-relevant uses form the majority in every source, but the balance of R-Labels varies. Except in the seven-instance Halmos sample, receiving no R-Label is the most frequent outcome in the R dimension

²Data, annotations, code, and evaluation results are available at <https://github.com/kmineshima/but-in-mathematical-proofs>.

Source	Corpus		P-Label			R-Label			
	Proofs	Med. length	Proof	Proof-neg	Meta	Contrast	Exception	No R-Label	but occ.
Velleman	39	118	36	8	2	7	0	39	46
Enderton (set theory)	65	161	37	19	18	21	5	48	74
Enderton (logic)	46	123	25	8	13	9	2	35	46
Landman	43	98	44	13	4	23	2	36	61
Partee et al.	26	99.5	15	10	2	4	0	23	27
Halmos	6	98.5	6	0	1	5	1	1	7
Total	225	124	163	58	40	69	10	182	261

Table 2: Corpus size, median proof length, and final-label distribution by source. Length is measured in word-like tokens. The **Proofs** and **Med. length** columns summarize the 225 unique proofs; label columns and *but* occurrences use 261 target instances.

for every source; the unusually high proportion of CONTRAST in Halmos should therefore be interpreted cautiously. These source-level differences are descriptive and do not by themselves establish effects of mathematical domain or authorial style.

Proof-advancing functions. The PROOF and PROOF-NEG labels mark proof-advancing functions: in both cases, the material associated with *but* contributes to the derivation as a whole. They may invoke an assumption, definition, or earlier result; open a case; or state a consequence completing the current step. We do not annotate finer transition types, whose correspondence to formal operations depends on the chosen formalization. Moreover, *but* cues this proof-advancing function but does not itself make the associated content proof-relevant: the same contribution could be unmarked or differently marked, and deleting only *but* need not make the proof incomplete. The 158 P-only cases show that proof-advancing material frequently occurs without an identifiable opposition or exception.

Supplementary analyses of sentence position and recurring compound expressions are provided in Appendix B.

4.2 Qualitative Analysis and Discussion

To examine internal variation, we focus on recurrent patterns within the PROOF and META labels.

4.2.1 PROOF Label

The PROOF label covers several inferential forms. The examples below include both P-only cases and cases that also receive an R-Label. For example, in (10), *but* marks the point where the assumption $\forall a \in A(f(a) = g(a))$ is applied, separating the general derivation from the step that invokes the given premise.

(10) **Label:** PROOF

Suppose $\forall a \in A(f(a) = g(a))$, and let (a, b) be an arbitrary element of f . Then $b = f(a)$. **But** by our assumption $f(a) = g(a)$, so $b = g(a)$ and therefore $(a, b) \in g$. (Velleman, 2019, Proof of Theorem 5.1.4, p. 232)

Introducing assumptions. In (11), *but* contrasts two possibilities and introduces an alternative assumption that guides the subsequent reasoning.

(11) **Label:** PROOF, CONTRAST

If nothing at all is in $\mathcal{C}$, we are done, **but** suppose some function ψ is in $\mathcal{C}$. (Enderton, 2001, Proof of Theorem 36J, p. 260)

This use is equivalent to an implicit *otherwise*, marking a branching point where a new assumption is introduced.

Gap emphasis. Some occurrences of *but* emphasize a gap between a stated conclusion and its supporting reasoning.

In (12), $x \subseteq y$ is short for $x = y$ or $x \in y$. Thus, the statement “ $\alpha + \delta = \beta$ or $\alpha + \delta \in \beta$ ” appears before the *but*-clause, while the equality $\alpha + \delta = \beta$ is asserted after the clause.

(12) **Label:** PROOF, CONTRAST

[...] we know from Theorem 8D that there is a largest δ for which $\alpha + \delta \subseteq \beta$. **But** equality must hold, for [...]. (Enderton, 1977, p. 235)

There is a certain contrast between “ $x \subseteq y$ ” and “ $x = y$ ”, and the proof gap is filled by the subsequent clause beginning with *for*. This usage can be regarded as marking a shift to *backward reasoning*, in which the conclusion is stated first, followed by the supporting grounds, rather than proceeding from premises to conclusion.

Additional examples of these qualitative patterns are provided in Appendix C.

4.2.2 META Label

The META label includes expository material that helps readers follow a proof without contributing a necessary inference. One recurrent subtype presents the current proof goal, typically with expressions such as “we must prove” or “we claim.”

(13) **Label:** META, CONTRAST

Obviously, *at most* one of the alternatives $x \subset y$, $x = y$, or $y \subset x$ can hold, **but** we must prove that at least one holds.

(Enderton, 1977, Proof of Theorem 5RA, pp. 113–114)

Here, *but* contrasts “at most one” with the remaining goal of proving “at least one.” Because stating that goal guides the exposition but is not itself required for the derivation, the occurrence is labeled META, CONTRAST. Additional META examples are provided in Appendix D.

4.3 Illustrative Comparison with Formal Proof Structure

We present three purposively selected case studies from Velleman (2019), using corresponding code provided by Velleman (2023) for the Lean proof assistant (Avigad and Massot, 2025): two relatively direct alignments in Appendix E and one mismatch below. The cases illustrate possible correspondence patterns rather than their prevalence. Because the mapping depends on the chosen formalization, they clarify why the P-Label criterion cannot require line-level correspondence.

- (14) **Theorem.** Suppose B is a set and F is a family of sets. If $\bigcup F \subseteq B$ then $F \subseteq \mathcal{P}(B)$. (Velleman, 2019, p. 125)

Proof. (Velleman, 2019, p. 125)

- (L1) Suppose $\bigcup F \subseteq B$.
- (L2) Let x be an arbitrary element of F .
- (L3) Let y be an arbitrary element of x .
- (L4) Since $y \in x$ and $x \in F$, by the definition of $\bigcup F$, $y \in \bigcup F$.
- (L5) **But** then since $\bigcup F \subseteq B$, $y \in B$.
- (L6) Since y was an arbitrary element of x , we can conclude that $x \subseteq B$, so $x \in \mathcal{P}(B)$.
- (L7) **But** x was an arbitrary element of F , so this shows that $F \subseteq \mathcal{P}(B)$, as required.

The following is the corresponding Lean proof given by Velleman (2023, pp. 40–43).

```
1 theorem Example_3_3_5 (U : Type) (B : Set U)
2 (F : Set (Set U)) :  $\bigcup_0 F \subseteq B \rightarrow F \subseteq \mathcal{P} B$  := by
3   assume h1 :  $\bigcup_0 F \subseteq B$ 
4   define
5     fix x : Set U
6     assume h2 :  $x \in F$ 
7     define
8       fix y : U
9       assume h3 :  $y \in x$ 
10    define at h1
11      apply h1 _
12    define
13      apply Exists.intro x _
14    show  $x \in F \wedge y \in x$  from And.intro h2 h3
15    done
```

The informal proof explicitly marks universal-generalization steps in (L6) and (L7), mainly for pedagogical reasons, while no single Lean line directly corresponds to the *but*-clause in (L5). The formal proof instead completes the same inferential work from its goal and local context. Thus, a *but*-clause can be proof-relevant even without a one-to-one formal counterpart: the P-Label is determined by its role in the informal derivation, not by surface alignment with Lean code.

5 Automatic Label Prediction

5.1 Experimental Setup

We evaluated label prediction using grouped stratified five-fold cross-validation, pooling held-out predictions across folds. Proofs with the same normalized statement were assigned to the same fold, yielding 212 split groups from 225 proof records (261 instances) and preventing closely related proofs from crossing the train–test boundary.

Baselines were the training-fold majority and a class-balanced linear softmax classifier using sentence-level word unigram and bigram TF–IDF features. The linear classifier receives the target sentence with the target occurrence replaced by a special marker, and a separate classifier is trained for each task. Through OpenRouter, we also evaluated GPT-5.6 Terra and the open-weight Qwen3-8B (Qwen Team, 2025), comparing marked target-sentence and full-proof input under zero- and eight-shot prompting. The eight demonstrations represented the eight observed combined labels and were selected deterministically from training-fold instances; none came from a held-out split group. Appendix F gives the full settings; data, predictions, and code are available in the public repository linked in Section 4.1.

For the P-Label, R-Label, and complete hierarchical label, we report accuracy and macro-F1,

Model	Input	Prompt	P-Label		R-Label		Hierarchical	
			Acc.	F1	Acc.	F1	Acc.	F1
Majority	Sentence	—	.625	.256	.697	.274	.448	.077
TF-IDF linear	Sentence	—	.667	.570	.797	.716	.579	.476
Terra	Sentence	Zero	.751	.708	.747	.569	.586	.413
Terra	Sentence	8-shot	.751	.695	.720	.700	.556	.379
Terra	Full proof	Zero	.759	.724	.628	.615	.498	.363
Terra	Full proof	8-shot	.739	.705	.655	.677	.494	.348
Qwen3-8B	Sentence	Zero	.460	.333	.655	.422	.341	.137
Qwen3-8B	Sentence	8-shot	.682	.574	.602	.559	.456	.314
Qwen3-8B	Full proof	Zero	.241	.143	.318	.200	.080	.029
Qwen3-8B	Full proof	8-shot	.670	.499	.421	.428	.303	.217

Table 3: Pooled held-out results over five grouped folds. F1 is macro-F1. Bold indicates the best value in each metric column.

which gives equal weight to rare labels. For sentence/full-proof accuracy differences, unadjusted percentile 95% CIs were obtained from 20,000 paired bootstrap resamples of the 212 groups; we make no dichotomous significance claims.

5.2 Results and Error Analysis

Table 3 shows that sentence-only zero-shot Terra achieves the highest hierarchical accuracy (.586), only 0.7 points above the TF-IDF baseline (.579), whose macro-F1 is higher (.476 versus .413). Terra therefore does not consistently outperform the linear baseline, especially on rare labels. Eight-shot prompting raises Terra’s sentence-level R-Label macro-F1 (.569 to .700) but lowers hierarchical accuracy; for Qwen, it broadens the label set and raises hierarchical accuracy by 11.5 points with sentence input and 22.2 points with full proofs.

In the fixed runs, full-proof input yielded lower hierarchical accuracy in all four comparisons. The full-proof-minus-sentence differences were -8.8 points (95% CI $[-16.0, -1.5]$) and -6.1 points ($[-12.5, 0.4]$) for Terra, and -26.1 points ($[-32.4, -19.7]$) and -15.3 points ($[-21.7, -8.9]$) for Qwen in the zero- and eight-shot conditions, respectively.

Prediction distributions suggest that additional context introduced competing cues despite explicit target markers. Terra zero-shot assigns CONTRAST 84 times from sentences but 149 times from full proofs (69 gold), while Qwen’s zero-shot full-proof run assigns PROOF-NEG, CONTRAST to 239 of 261 instances. Among the paired Terra zero-shot predictions, 59 instances are correct only with sentence input, compared with 36 only with full-proof input. For example, for “But $a \notin B \setminus C$, so it follows that

$a \in C$,” the gold and sentence-only Terra labels are PROOF, whereas the full-proof prediction adds CONTRAST.

Errors also track annotation difficulty: the best Terra condition reaches .670 hierarchical accuracy on the 176 instances with identical pre-adjudication labels, but .412 on the 85 disagreement instances. Frequent errors add CONTRAST to PROOF or confuse PROOF with PROOF-NEG. Exceptive combinations remain difficult: sentence-only zero-shot Terra predicts Exception only four times (10 gold) and never correctly predicts META, EXCEPTION; this helps explain the low hierarchical macro-F1 despite stronger P-Label accuracy.

6 Summary and Prospects

This study constructed and annotated a focused corpus of 261 occurrences of *but* in 225 informal proofs from six textbooks. Its hierarchical scheme separates proof-level contribution from optional contrastive and exceptive uses, treating connective use as part of mathematical reasoning. Of the 261 occurrences, 221 (84.7%) marked proof-relevant transitions. Qualitative analysis and three Lean comparisons further showed that proof relevance cannot be reduced to line-level correspondence with formal proofs. Prediction experiments established initial benchmarks; in the fixed runs, full-proof input yielded lower accuracy in all four comparisons, with Qwen3-8B decreases of 15.3–26.1 points. Together, the corpus, scheme, and analyses provide an empirical basis for studying connectives in informal mathematical language. Future work should extend the analysis to other connectives and implicit relations and examine informal-formal correspondence on larger datasets.

Limitations

The corpus is small and domain-specific: it contains 261 occurrences of *but* from six textbooks and excludes other connectives and implicit discourse relations. The results should therefore be read as a case study of explicit *but*, not as a general inventory or frequency comparison of discourse relations in mathematical proofs. Comparisons with non-mathematical expository discourse are needed before treating the observed P-only patterns as specific to mathematical proof.

The annotation scheme necessarily involves interpretive judgments, particularly for R-Labels and the functionally heterogeneous P-only cases. Resolved disagreements concentrated at the boundary between PROOF and META, the distinction between intermediate proof steps and completed contradictions, and the decision whether an explicit R-Label was warranted. The EXCEPTION subtype remains provisional: only 10 instances received this adjudicated label, and both independent annotators selected it in only five of them.

Both supplementary evaluations are exploratory. The Lean analysis consists of three purposively selected cases—two relatively direct alignments and one mismatch—and illustrates possible correspondence patterns rather than their prevalence. The mapping also depends on the chosen formalization. The prediction benchmark is limited to two LLMs, one linear baseline, two context windows, and one fixed zero-shot or eight-shot prompt per condition, without repeated stochastic runs, prompt optimization, or fine-tuning. Label imbalance makes macro-F1 uncertain. The bootstrap intervals reflect variation across the sampled proof groups, not repeated model runs, alternative prompts, or providers; the observed full-proof differences may therefore depend on prompting.

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions, which have improved this paper. This work was partially supported by JST CREST Grant Number JPMJCR2114, JST BOOST Grant Number JPMJBS2409 and JSPS KAKENHI Grant Number JP24K03340, JP24K00004.

During manuscript preparation, we used OpenAI’s Codex to assist with language editing and checks of internal consistency and bibliographic information. The authors reviewed the resulting

suggestions and take full responsibility for the final content.

References

- Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.
- Andrea Asperti, Alberto Naibo, and Claudio Sacerdoti Coen. 2026. Thinking machines: Mathematical reasoning in the age of LLMs. *Big Data and Cognitive Computing*, 10(1):38.
- Jeremy Avigad and Patrick Massot. 2025. Mathematics in Lean. Release v4.19.0. https://avigad.github.io/mathematics_in_lean/mathematics_in_lean.pdf.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.
- Marcos Cramer, Bernhard Fisseni, Peter Koepke, Daniel Kühlwein, Bernhard Schröder, and Jip Veldman. 2009. The Naproche project controlled natural language proof checking of mathematical texts. In *International Workshop on Controlled Natural Language*, pages 170–186. Springer.
- Herbert B Enderton. 1977. *Elements of Set Theory*. Academic Press.
- Herbert B Enderton. 2001. *A Mathematical Introduction to Logic*, second edition. Academic Press.
- Mohan Ganesalingam. 2013. *The Language of Mathematics: A Linguistic and Philosophical Investigation*. Springer.
- Paul R. Halmos. 1974. *Lectures on Boolean Algebras*. Springer.
- Rodney D. Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.
- Cezary Kaliszyk, Josef Urban, and Jiří Vyskočil. 2017. Automating formalization by statistical and semantic parsing of mathematics. In *International Conference on Interactive Theorem Proving*, pages 12–27. Springer.
- Andrew Kehler. 2002. *Coherence, Reference, and the Theory of Grammar*. CSLI publications.
- Alexander W. Kocurek and Carlotta Pavese. 2022. The dynamics of argumentative discourse. *Journal of Philosophical Logic*, 51:413–456.
- Fred Landman. 1991. *Structures for Semantics*. Springer.
- Takuya Matsuzaki, Takumi Ito, Hidenao Iwane, Hirokazu Anai, and Noriko H. Arai. 2017. Semantic parsing of pre-university math problems. In *Proceedings of the 55th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 2131–2141, Vancouver, Canada. Association for Computational Linguistics.

Barbara H. Partee, Alice ter Meulen, and Robert E. Wall. 1990. *Mathematical Methods in Linguistics*. Springer.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics*, 40(4):921–950.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Aarne Ranta. 1995. Syntactic categories in the language of mathematics. In *Types for Proofs and Programs*, pages 162–182. Springer.

Christian Szegedy. 2020. A promising path towards autoformalization and general artificial intelligence. In *International Conference on Intelligent Computer Mathematics*, pages 3–20. Springer.

Fenner Stanley Tanswell and Matthew Inglis. 2024. The language of proofs: A philosophical corpus linguistics study of instructions and imperatives in mathematical texts. In Bharath Sriraman, editor, *Handbook of the History and Philosophy of Mathematical Practice*, pages 2925–2952. Springer.

Daniel Velleman. 2023. How to prove it with Lean. URL: <https://djvvelleman.github.io/HTPIwL>.

Daniel J. Velleman. 2019. *How to Prove It: A Structured Approach*, third edition. Cambridge University Press.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn discourse treebank 3.0 annotation manual. Technical report, University of Pennsylvania.

Ke Weng, Lun Du, Sirui Li, Wangyue Lu, Haozhe Sun, Hengyu Liu, and Tiancheng Zhang. 2025. Autoformalization in the era of large language models: A survey. *arXiv preprint arXiv:2505.23486*.

Grégoire Winterstein. 2012. What *but*-sentences argue for: An argumentative analysis of *but*. *Lingua*, 122(15):1864–1885.

Sandrine Zufferey and Liesbeth Degand. 2024. *Connectives and Discourse Relations*. Cambridge University Press.

A Annotation Guidelines

This appendix consolidates the decision rules used for the hierarchical annotation. The target of annotation is one overt token of *but* or *But*; annotators inspect the complete proof containing that token, not only the sentence in which it occurs.

Step 1: assign the P-Label. The P-Label concerns the content associated with *but*, not whether the connective itself can be deleted. The two proof-relevant outcomes, PROOF and PROOF-NEG, jointly operationalize proof-advancing transitions at a coarse level; META does not. Except for distinguishing contradiction-completing transitions as PROOF-NEG, the annotation does not assign a more specific transition type. Annotators apply the following decision order.

1. If removing the associated content leaves the stated propositions and inference steps sufficient for the proof, assign META. Typical cases are terminological cautions, explanatory asides, and comments that improve readability without advancing the derivation.
2. Otherwise, if the content states or completes a contradiction used to reject an assumption or derive a negated conclusion, assign PROOF-NEG.
3. Otherwise, assign PROOF. This includes required assumptions, applications of definitions or earlier results, case distinctions, and consequences that complete a proof obligation.

Because PROOF-NEG is the more specific proof-relevant label, it takes precedence over PROOF whenever its contradiction criterion is satisfied.

Two boundary rules are important. First, mathematical content alone does not warrant PROOF: necessity is evaluated relative to the stated proof obligation, so goal or method announcements and continuations after that obligation has already been discharged may be META. Second, occurring within a reductio argument does not by itself warrant PROOF-NEG. The target content must state or complete the contradiction; an intermediate consequence that merely advances the reductio remains PROOF.

Step 2: assign an R-Label only if applicable. R-Labels are evaluated after the P-Label and independently of proof relevance; for example, both PROOF, CONTRAST and META, CONTRAST are possible. An R-Label is optional and is not an alternative to the P-Label: inferential continuation, goal presentation, or other proof-level functions do not preclude an R-Label when an explicit contrast or exception is also present.

1. Test EXCEPTION first. Assign it only if a generalization or shared frame remains applicable except for an explicitly excluded or substituted member, case, parameter, or respect, and the passage permits a meaning-preserving paraphrase with *except* or *except for*.
2. If the exceptive test fails, test CONTRAST. Assign it when the text explicitly aligns two entities, properties, states, or alternatives and highlights their opposition or difference. The aligned pair must be identifiable from the text; a clause’s inferential importance alone is insufficient. Denial-of-expectation and concessive readings are included under CONTRAST; no separate Concession label is assigned.
3. If neither test succeeds, assign no R-Label. Absence of an R-Label is not a third relation category.

The EXCEPTION label includes phrasal exceptive uses and therefore should not be interpreted as a PDTB-style discourse-relation label. Negation, contradiction, or the general adversative force of *but* is not by itself sufficient for CONTRAST; the aligned entities, properties, states, or alternatives must be textually identifiable.

Step 3: record the combined label. The P-Label is written first. A P-only output is written as PROOF; when an R-Label is assigned, the two labels are separated by a comma, as in PROOF, CONTRAST or META, EXCEPTION.

Supplementary fields. Annotators also record whether the token is sentence-initial or sentence-medial and note recurring compound expressions such as *but then*, *but since*, and *but when*. These fields do not determine the P-Label or R-Label.

Uncertainty and adjudication. Annotators completed the two label decisions independently and could record *uncertain* or *other* for difficult cases. These were response options for flagging problems, not labels in the final scheme. After both annotations were complete, all disagreements and flagged cases were discussed by the six annotators; every instance received an adjudicated P-Label and a final decision among CONTRAST, EXCEPTION, and no R-Label. All 261 adjudicated instances were retained in the distributional analyses and prediction experiments.

Agreement calculation. The primary agreement analysis retains *uncertain* and *other* as distinct responses, thereby including rather than selectively removing difficult cases. As a sensitivity analysis, we also excluded any case in which either annotator used one of these responses. P-Label agreement then increased from 210/261 (80.5%; Fleiss’ $\kappa = .615$) to 210/254 (82.7%; $\kappa = .647$), while complete hierarchical-label agreement increased from 176/261 (67.4%; $\kappa = .554$) to 176/253 (69.6%; $\kappa = .578$). The numerators are unchanged because all excluded cases were disagreements, and the overall pattern is stable under the alternative treatment.

R-Label reliability. We evaluated R decisions separately after excluding four cases in which either annotator gave an unresolved R response. Agreement was 220/257 (85.6%; $\kappa = .661$) for the binary decision between no R-Label and an assigned R-Label, and 215/257 (83.7%; $\kappa = .636$) for the exact three-way decision among CONTRAST, EXCEPTION, and no R-Label. The two annotators gave matching valid P-Labels in 210 cases; after excluding the one case with an unresolved R response, exact R agreement was 176/209 (84.2%; $\kappa = .643$).

R1 \ R2	Contrast	Exception	No R	Total
CONTRAST	50	4	14	68
EXCEPTION	1	5	9	15
No R-Label	14	0	160	174
Total	65	9	183	257

Table 4: Independent R-Label decisions for the 257 cases with resolved R responses.

Because annotator pairs varied across items, the rows and columns in Table 4 denote annotation slots rather than fixed individuals. When both annotators assigned a subtype, they agreed on CONTRAST versus EXCEPTION in 55/60 cases (91.7%; $\kappa = .619$); the five remaining cases were direct subtype confusions.

Adjudicated EXCEPTION cases. The 10 adjudicated EXCEPTION cases show greater uncertainty than this aggregate rate suggests. Their independent R pairs were EXCEPTION / EXCEPTION (5 cases) and EXCEPTION / no R-Label (5). Thus both annotators independently selected EXCEPTION in 5/10 cases, exactly one did so in 5/10, and neither did so in 0/10. Because this subset is defined by the adjudicated outcome, these counts are treated

as a descriptive audit rather than as the basis for a separate κ .

Disagreement patterns. Among the 44 P-Label disagreements with resolved responses, PROOF / META and PROOF / PROOF-NEG each occurred 21 times, while META / PROOF-NEG occurred twice. Adjudication selected META in 15 of the 21 PROOF / META cases and PROOF-NEG in 17 of the 21 PROOF / PROOF-NEG cases. Eight of the latter 21 cases contained *but then*, which often advances a contradiction argument without determining whether the target sentence itself completes the contradiction.

The following paired examples illustrate why full proof context and mathematical knowledge are needed at the two most frequent P-Label boundaries.

- (15) **Independent P-Labels:** META / PROOF
Adjudicated label: META
 Since α is a set, [...] some ordinals do not belong to α . This proves the theorem, **but** we will continue and show that α itself is the least such ordinal. (Enderton, 1977, Proof of Hartogs' Theorem, p. 196)
- (16) **Independent P-Labels:** PROOF / META
Adjudicated label: PROOF
 Assume that $g \circ f$ is not onto. [...] **But** g is onto C , and thus there is some $y \in B$ such that $g(y) = z$. (Partee et al., 1990, Exercise 6, p. 581)

In (15), the stated theorem has already been established, so the following strengthening is META. In (16), surjectivity supplies a witness required for the subsequent contradiction, so the target is PROOF.

- (17) **Independent P-Labels:** PROOF / PROOF-NEG
Adjudicated label: PROOF
 Thus if A is recursive, then f (as a relation) is recursively enumerable. **But** then f is a total recursive function, and so equals $\llbracket e \rrbracket_1$ for some $e \in A$. (Enderton, 2001, Proof of Theorem 36E, p. 254)
- (18) **Independent P-Labels:** PROOF / PROOF-NEG
Adjudicated label: PROOF-NEG
 [...] the only other possibility is $x = b$. **But** then since xRc we must have bRc , which contradicts our assumption that

$\neg bRc$. (Velleman, 2019, Solution to Example 6.2.1, p. 346)

Example (17) is an intermediate consequence; the contradiction appears only in the next sentence. Example (18) explicitly completes the contradiction and is therefore PROOF-NEG.

Table 4 shows that the R-applicability boundary produced 37 disagreements, compared with five direct CONTRAST / EXCEPTION confusions. Of the 37 disagreements over whether to assign any R-Label, adjudication assigned one in 18 cases and none in 19, confirming a genuine boundary. Example (2) illustrates no R-Label: its negative premise advances the proof without aligning opposing alternatives. By contrast, Example (19) aligns properties of the same object, while Example (20) preserves a proof frame but replaces one operation with another.

- (19) **Independent R decisions:**
 CONTRAST / no R-Label
Adjudicated label: PROOF, CONTRAST
 Then there exists some $t \in A$ with $\text{seg } t \subseteq B$ **but** $t \notin B$. (Enderton, 1977, Proof of Theorem 7C, p. 175)
- (20) **Independent R decisions:**
 EXCEPTION / no R-Label
Adjudicated label: PROOF, EXCEPTION
 The proof is exactly as in Lemma 5ZB, **but** with addition replaced by multiplication. (Enderton, 1977, Proof of Lemma 5QD, p. 106)

B Sentence Position and Compound Expressions

Sentence position appears to interact systematically with discourse function. As summarized in Table 5, sentence-initial *But* appeared in 173 out of 261 cases (66.3%), predominantly within proof-relevant contexts (PROOF=103, PROOF-NEG=50). In contrast, sentence-medial *but* occurred mostly in PROOF contexts, though a smaller proportion appeared in META clauses.

Table 6 shows that sentence-initial *But* frequently co-occurred with the absence of an R-Label (150/173 = 86.7%), whereas sentence-medial *but* carried CONTRAST or EXCEPTION in 56 of 88 cases (63.6%). This pattern is consistent with initial *But* serving as a discourse-level transition cue and medial *but* expressing local contrast or exception within proof steps.

Among compound expressions, three combinations were notable: (i) *but then* appeared most frequently (52 instances), all in proof-relevant contexts (PROOF: 33; PROOF-NEG: 19), and received no R-Label in 47 cases; (ii) *but since* occurred 16 times (PROOF: 13; PROOF-NEG: 3) and received no R-Label in 15 cases, usually introducing a step justified by a stated reason; and (iii) *but when* was rare (2 instances), both labeled META with no R-Label.

Label	Initial	Medial	Total
PROOF	103	60	163
PROOF-NEG	50	8	58
META	20	20	40
Total	173	88	261

Table 5: Proof-relevance label by sentence position. Initial denotes sentence-initial; medial denotes sentence-medial.

R-Label	Initial	Medial	Total
CONTRAST	23	46	69
EXCEPTION	0	10	10
No R-Label	150	32	182
Total	173	88	261

Table 6: Optional relational-use subtype by sentence position. Initial denotes sentence-initial; medial denotes sentence-medial.

C Additional Qualitative Examples for the PROOF Label

This appendix supplements the three representative examples retained in Section 4.2.

Applying an inductive hypothesis. In (21), *but* introduces the application of the inductive hypothesis after a step based on the definition of B_{n+1} . It thus cues a shift between adjacent inferential operations rather than an explicit contrast.

(21) Label: PROOF

Now suppose $x \in B_{n+1}$. By the definition of B_{n+1} this means $x \in f(B_n)$, so there is some $b \in B_n$ such that $x = f(b)$. **But** by the inductive hypothesis, $B_n \subseteq D$, so $b \in D$, and since D is closed under f it follows that $x = f(b) \in D$. (Velleman, 2019, Proof of Theorem 6.5.1, p. 318)

In (22), *but* introduces the decisive application of the one-to-one property of f , moving from $f(a_1) = f(a_2)$ to $a_1 = a_2$.

(22) Label: PROOF

Thus, $a_2 \notin Y$, so $a_2 \in X$. This means that $h(a_2) = f(a_2)$, so from the equation $h(a_1) = h(a_2)$ we get $f(a_1) = f(a_2)$. **But** f is one-to-one, so it follows that $a_1 = a_2$. (Velleman, 2019, Proof of Theorem 8.3.2, p. 392)

Introducing assumptions. As in the example retained in the main text, *but* in (23) marks a branching point and introduces the alternative assumption that S is nonempty.

(23) Label: PROOF, CONTRAST

We hope that S is empty. **But** if not, then there is a least $t \in S$.

(Enderton, 1977, Proof of Theorem 7D, p. 183)

Trivial steps. Expressions such as “But this is clear” may introduce an immediate but logically indispensable step and are therefore assigned PROOF.

(24) Label: PROOF

We must show that $\text{card}(V_\omega) = \aleph_0$. **But** this is clear, since V_ω is the union of $\aleph_0$ finite sets of increasing size.

(Enderton, 1977, Proof of Theorem 9J, p. 252)

Gap emphasis. In (25), the *but*-clause derives $\text{card } S \geq \aleph_{\alpha+1}$ from $\aleph_{\alpha+1} \leq (\text{card } S) \cdot \aleph_\alpha$ while leaving the relevant cardinal-arithmetic properties implicit. The connective draws attention to an inferential step for which further explanation could be supplied.

(25) Label: PROOF

Therefore [...] we have $\aleph_{\alpha+1} = \text{card } \bigcup S \leq (\text{card } S) \cdot \aleph_\alpha$. **But** this implies that $\text{card } S \geq \aleph_{\alpha+1}$. (Enderton, 1977, Proof of Theorem 9M, p. 256)

D Subclasses of META Label

This appendix describes subclasses of the META label. These cases typically involve *but* appearing in remarks, qualifications, or side comments within proofs. Although omitting these segments would not affect the validity of the proof, they contribute to clarity and readability and are therefore annotated as META.

Adding remarks. In this subclass, *but* introduces a comment or clarification outside the core inferential steps of the proof.

(26) **Label:** META, CONTRAST

Clearly f is onto, since by definition all rational numbers can be written as a fraction, **but** note that f is not one-to-one.

(Velleman, 2019, Proof of Theorem 8.1.6, p. 458)

(27) **Label:** META

The methods of Sections 3.3 and 3.4 can be applied to the Gödel numbers of these formulas to show that p is recursive. **But** the details are not particularly attractive, and we omit them here.

(Enderton, 2001, Proof of Lemma 37E, p. 273)

Interrogative goal presentation. The following example illustrates the use of an interrogative within a proof, used in presenting the goal of the argument. The presence of such non-declarative sentences demonstrates the diversity of linguistic forms that appear in informal proofs.³

(28) **Label:** META

We must show that $\phi(\alpha, y) \Leftrightarrow \gamma(f, y)$ for each y . **But** what is f ? It must be the function F_α , by the above.

(Enderton, 1977, pp. 210–212)

E Detailed Lean Comparisons

This appendix presents the two relatively direct alignment cases that complement the mismatch case in Section 4.3. We use the Lean library and tactics provided by Velleman (2023).

A PROOF-NEG example. Consider the following theorem and informal proof.

(29) **Theorem.** Suppose A , B , and C are sets, $A \setminus B \subseteq C$, and x is anything at all. If $x \in A \setminus C$ then $x \in B$. (Velleman, 2019, p. 107)

Proof. (Velleman, 2019, p. 107)

(L1) Suppose $x \in A \setminus C$.

(L2) This means that $x \in A$ and $x \notin C$.

(L3) Suppose $x \notin B$.

³For a corpus-based analysis of imperative expressions in proofs, such as *Suppose that* and *Let*, see Tanswell and Inglis (2024).

(L4) Then $x \in A \setminus B$, so since $A \setminus B \subseteq C$, $x \in C$.

(L5) **But** this contradicts the fact that $x \notin C$.

(L6) Therefore $x \in B$.

(L7) Thus, if $x \in A \setminus C$ then $x \in B$.

The corresponding Lean proof is:

```

1 theorem Example_3_2_3
2   (U : Type) (A B C : Set U) (x : U)
3   (h : A \ B \subseteq C)
4   (h1 : x \in A \setminus C) :
5   x \in B := by
6   define at h1
7   by_contra h2
8   have h3 : x \in A := h1.left
9   have h4 : x \in A \wedge x \notin B := And.intro h3 h2
10  define at h
11  have h5 : x \in C := h h4
12  show False from h1.right h5
13 done

```

The theorem declaration contains the assumptions and conclusion, with (L1) represented by the hypothesis `h1`; the next command expands this hypothesis as in (L2). The command `by_contra h2` introduces the negated conclusion corresponding to (L3), after which the proof derives $x \in A \setminus B$ and then $x \in C$, matching (L4). Finally, show `False` from `h1.right h5` corresponds to the contradiction stated in the *but*-clause in (L5); the remaining informal sentences close the argument.

A PROOF example. The next case also permits a relatively direct comparison.

(30) **Theorem.** There is a unique set A such that for every set B , $A \cup B = B$. (Velleman, 2019, p. 157)

Proof. (Velleman, 2019, pp. 157–158)

(L1) Existence: Clearly $\forall B(\emptyset \cup B = B)$, so $\emptyset$ has the required property.

(L2) Uniqueness: Suppose $\forall B(C \cup B = B)$ and $\forall B(D \cup B = B)$.

(L3) Applying the first assumption to D gives $C \cup D = D$, and applying the second to C gives $D \cup C = C$.

(L4) **But** clearly $C \cup D = D \cup C$, so $C = D$.

The corresponding Lean proof is (Velleman, 2023, pp. 83–87):

```

1 theorem Example_3_6_2 (U : Type) :
2   \exists! (A : Set U), \forall (B : Set U), A \cup B = B := by
3   exists_unique
4   -- Existence
5   apply Exists.intro \emptyset
6   show \forall (B : Set U), \emptyset \cup B = B from empty_union
7   done
8   -- Uniqueness
9   fix C : Set U; fix D : Set U
10  assume h1 : \forall (B : Set U), C \cup B = B
11  assume h2 : \forall (B : Set U), D \cup B = B
12  have h3 : C \cup D = D := h1 D
13  have h4 : D \cup C = C := h2 C

```

```

14 show C = D from
15   calc C
16   _ = D ∪ C := h4.symm
17   _ = C ∪ D := union_comm D C
18   _ = D := h3
19 done
20 done

```

The `exists_unique` tactic divides the proof into existence and uniqueness: the existence block corresponds to (L1), while the assumptions and their applications in the uniqueness block correspond to (L2) and (L3). The final `calc` block uses commutativity of union to conclude $C = D$, matching the inferential role of the *but*-clause in (L4).

F Model Evaluation Reproducibility

Models and inference settings. Both LLMs were queried through the OpenRouter Chat Completions API with JSON Schema output restricted to the eight observed combined labels in Table 1. For Terra, the model identifier was `openai/gpt-5.6-terra`; requests were pinned to Azure with provider fallbacks disabled, reasoning effort set to low, temperature left at the provider default, and `max_completion_tokens=256`. For Qwen3-8B (Qwen Team, 2025), the identifier was `qwen/qwen3-8b`; all response metadata identified the Alibaba FP8 endpoint, thinking was disabled, temperature was 0, and `max_tokens=64`. Each condition was run once.

Input construction. The sentence condition supplied only the sentence containing the target token, whereas the full-proof condition supplied the complete proof but not its statement. In both conditions, the relevant token was enclosed in `<TARGET>` and `</TARGET>` tags. No input truncation was applied: full proofs and all eight demonstrations were passed in full. The longest serialized prompt contained 14,137 characters.

The artifact does not redistribute the source statements or complete proof texts because they originate in copyrighted textbooks; it provides the annotations, split metadata, completed predictions, and analysis code needed to audit the reported results.

Complete system prompt. The following reproduces the static system prompt; only typography has been normalized for print.

You annotate the discourse function of one marked occurrence of “but” in an informal mathematical proof.

Apply the hierarchy in this order.

Step 1—P-label:

- **Meta:** removing the associated content leaves the stated propositions and inference steps sufficient. Typical cases are terminological cautions, explanatory asides, and readability comments.
- **Proof-neg:** the content states or completes a contradiction used to reject an assumption or derive a negated conclusion.
- **Proof:** the content is otherwise required for the derivation, including assumptions, definitions, earlier results, case distinctions, and consequences completing a proof obligation.

Proof-neg takes precedence over Proof when its criterion holds.

Step 2—optional R-label, independently of the P-label:

- **Exception:** test first. A generalization or shared frame remains applicable except for an explicitly excluded or substituted member, case, parameter, or respect; an “except” or “except for” paraphrase preserves the meaning.
- **Contrast:** if Exception fails, assign Contrast only when the text explicitly aligns identifiable entities, properties, states, or alternatives and highlights an opposition or difference.
- Otherwise assign no R-label.

The target token is enclosed in `<TARGET>` and `</TARGET>`. Classify the associated content, not whether the word “but” itself can be deleted.

User-prompt templates. The zero-shot user message was “Classify this target: [marked input]. Apply Step 1 and Step 2, then return the combined label.” Line breaks separated the introductory text, marked input, and final instruction. In the eight-shot condition, the message began “Here are labeled examples drawn only from the training portion of this fold:” and then repeated the block “Example: [marked demonstration]. Label: [combined label]” eight times before the zero-shot target block. The structured output schema permitted the following eight labels, in order: PROOF; PROOF, CONTRAST; PROOF, EXCEPTION; PROOF-NEG; PROOF-NEG, CONTRAST; META; META, CONTRAST; and META, EXCEPTION.

For each fold and combined label, one demonstration was selected deterministically from the training instances using a SHA-256 ordering over the fold, label, and instance identifier; no demonstration belonged to a held-out split group. The grouped five-fold split was generated deterministically and kept fixed across all experimental conditions, and deterministic per-request seeds were supplied to OpenRouter. The paired group-bootstrap analysis used 20,000 resamples and seed 20260730.