Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 11, 2026: 4822-4836
ISSN:2302-8149
e-ISSN:2540-9719

Abstention berbasis *Threshold* pada Verifikasi Klaim *Hoaks* berbahasa Indonesia menggunakan *IndoNLI*

Threshold-based Abstention for Indonesian Hoax Claim Verification using IndoNLI

[1]Arrizal Cahyadi, [2]Harits Ar Rosyid*, [3]M. Zainal Arifin
[1,2,3]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Negeri Malang
[1,2,3]Jl. Semarang No. 5, Malang, Jawa Timur 65145, Indonesia
*e-mail: harits.ar.ft@um.ac.id

(***received*:** 7 November 2026, ***revised*:** 9 November 2026, ***accepted*:** 11 November 2026)

Abstrak

Sistem verifikasi klaim hoaks otomatis perlu menahan diri ketika bukti kurang meyakinkan, tetapi model *natural language inference* (NLI) berbahasa Indonesia umumnya menjawab setiap masukan. Penelitian ini menguji apakah penolakan menjawab (*abstention*) berbasis *threshold* pada probabilitas *softmax* maksimum meningkatkan keandalan prediksi NLI. Model RoBERTa berbahasa Indonesia dilatih pada IndoNLI dengan tiga *seed* acak, dan *threshold* ditetapkan pada data validasi agar *coverage* sasaran tercapai. Evaluasi memakai IndoNLI *test_lay* dan *test_expert* serta 758 pasangan klaim dari TurnBackHoax (270 hoaks, label *contradiction*) dan Antara (488 berita valid, label *entailment*). Pada *coverage* sasaran 90%, akurasi pada data yang dijawab naik 4,32 ± 0,48 poin persentase pada *test_lay* (73,39% menjadi 77,71%) dan 3,60 ± 0,28 poin pada *test_expert* (60,27% menjadi 63,87%). Pada studi kasus, akurasi tanpa *abstention* hanya 62,14 ± 9,78% (55,41% sampai 73,35% antar *seed*), di bawah porsi kelas mayoritas (64,4%). Abstention menaikkan akurasi 7,36 ± 1,98 poin pada *coverage* nyata 71,5% dan menolak 49% sampai 65% hoaks yang diprediksi *entailment*, tetapi 39% sampai 65% hoaks yang salah prediksi tetap lolos. Pembanding *hypothesis-only* mencapai 59,3 ± 0,3% pada validasi IndoNLI (tebakan mayoritas 36,7%). Abstention berbasis *threshold* meningkatkan akurasi pada IndoNLI, tetapi belum cukup andal untuk dipakai langsung pada verifikasi klaim hoaks nyata. Label studi kasus berasal dari sumber artikel dan tidak memuat kelas *neutral*, sehingga hasilnya tidak digeneralisasi.

Kata kunci: *abstention, confidence threshold, hoax claim verification, IndoNLI, natural language inference*

Abstract

Automated hoax claim verification systems need to abstain from making predictions when the available evidence is insufficient or inconclusive. However, Indonesian-language Natural Language Inference (NLI) models generally produce predictions for every input. This study investigates whether threshold-based abstention using the maximum softmax probability can improve the reliability of NLI predictions. An Indonesian RoBERTa model was trained on the IndoNLI dataset using three random seeds, with the abstention threshold calibrated on the validation set to achieve a target coverage level. Evaluation was conducted on the IndoNLI test_lay and test_expert sets, as well as 758 claim pairs collected from TurnBackHoax (270 hoax claims labeled as contradiction) and Antara (488 valid news claims labeled as entailment). At a target coverage of 90%, accuracy among the retained predictions increased by 4.32 ± 0.48 percentage points on test_lay (from 73.39% to 77.71%) and by 3.60 ± 0.28 percentage points on test_expert (from 60.27% to 63.87%). In the case study, accuracy without abstention was only 62.14 ± 9.78% (ranging from 55.41% to 73.35% across seeds), which was below the majority-class baseline of 64.4%. Abstention increased accuracy by 7.36 ± 1.98 percentage points at an actual coverage of 71.5% and rejected 49% to 65% of hoax claims incorrectly predicted as entailment. However, 39% to 65% of misclassified hoax claims still passed through the system without being rejected. The hypothesis-only baseline achieved 59.3 ± 0.3% accuracy on the IndoNLI validation set, compared with 36.7% for majority-class prediction. These findings indicate that threshold-based abstention improves prediction accuracy on IndoNLI but remains insufficiently reliable for direct deployment in real-world hoax claim verification. The case

 ISSN:2302-8149
e-ISSN:2540-9719

study labels were derived from the source articles and did not include the neutral class; therefore, the findings should not be generalized beyond this evaluation setting.
Keywords: *abstention, confidence threshold, hoax claim verification, IndoNLI, natural language inference*

1 Pendahuluan

Hoaks beredar luas di ruang digital Indonesia. Kementerian Komunikasi dan Digital mengidentifikasi 1.923 konten hoaks sepanjang 2024 [1], sedangkan Masyarakat Anti Fitnah Indonesia (Mafindo) mencatat 1.593 kasus hoaks pada 21 Oktober 2024 sampai 17 Oktober 2025 [2]. Angka tersebut hanya mencakup konten yang berhasil teridentifikasi, sehingga pemeriksaan klaim secara otomatis dibutuhkan sebagai alat bantu bagi pemeriksa fakta [3].

Salah satu cara memeriksa klaim secara otomatis adalah *natural language inference* (NLI). Bukti diperlakukan sebagai premis dan klaim sebagai hipotesis, lalu model menentukan apakah bukti mendukung klaim (*entailment*), membantahnya (*contradiction*), atau tidak cukup untuk menyimpulkan (*neutral*). Pendekatan ini telah dipakai pada verifikasi fakta berbahasa Inggris [4] dan pada evaluasi konsistensi faktual [5]. Untuk bahasa Indonesia, IndoNLI [6] menyediakan hampir 18 ribu pasangan premis-hipotesis yang ditulis oleh anotator awam dan anotator ahli, sedangkan penelitian deteksi hoaks berbahasa Indonesia yang ditinjau mengklasifikasikan teks berita sebagai hoaks atau fakta [7], [8], [9], bukan memeriksa klaim terhadap bukti.

Kesesuaian NLI dengan pemeriksaan klaim terlihat dari rancangan dataset FEVER [10], yang memberi putusan didukung, dibantah, atau informasi tidak cukup untuk setiap klaim. Putusan ketiga sejajar dengan label *neutral* pada NLI. Berbeda dari klasifikasi artikel sebagai hoaks atau fakta, pemeriksa fakta bekerja pada tingkat klaim: mereka mencari bukti, lalu memutuskan apakah bukti itu cukup. Model bantu pemeriksa karena itu perlu tiga putusan yang sama, ditambah kemampuan menyatakan bahwa model sendiri tidak cukup yakin.

Masalahnya, model NLI selalu memilih satu dari tiga label, termasuk ketika bukti tidak memadai atau model tidak yakin. Model verifikasi fakta pun cenderung tetap memberi putusan pada bukti yang tidak lengkap [11]. Pada pemeriksaan hoaks, putusan keliru yang disampaikan dengan yakin dapat menyesatkan pengguna lebih jauh daripada tidak memberi jawaban. Penolakan menjawab (*abstention*) menjadi jalan keluar: sistem hanya menjawab jika cukup yakin dan menyerahkan sisanya kepada pemeriksa manusia [12], [13]. Bentuk paling sederhana adalah *threshold* pada probabilitas *softmax* tertinggi [14]. Bentuk ini bergantung pada kalibrasi probabilitas model, padahal jaringan saraf modern tidak selalu terkalibrasi [15], dan keunggulan metode prediksi selektif berbeda antar tugas dan domain [16]. Seberapa efektif *abstention* berbasis *threshold* pada NLI berbahasa Indonesia, dan apakah efeknya bertahan pada klaim studi kasus (TurnBackHoax dan Antara) yang berbeda dari data latih, belum terjawab oleh literatur yang ditinjau.

Evaluasi dengan satu kali pelatihan juga berisiko. Hasil *fine-tuning* dapat berbeda antar *seed* acak [17], dan variasi tersebut cukup besar untuk mengubah kesimpulan perbandingan model [18]. Penelitian ini bertujuan mengukur seberapa besar *abstention* berbasis *threshold* meningkatkan akurasi model NLI berbahasa Indonesia pada data IndoNLI, menguji apakah peningkatan itu bertahan pada 758 pasangan klaim (270 hoaks dan 488 berita valid) dari TurnBackHoax dan Antara, serta melaporkan variasi hasil pada tiga *seed* acak. Sebagai pembanding, penelitian ini juga mengukur seberapa jauh label dapat ditebak dari hipotesis saja. Threshold ditetapkan dari *coverage* sasaran pada data validasi untuk setiap *seed* agar perbandingan antar *seed* setara.

Penelitian ini dibatasi pada tahap penilaian, yaitu menentukan hubungan antara klaim dan bukti yang sudah tersedia. Pencarian bukti tidak termasuk karena premis sudah diberikan pada IndoNLI dan diambil dari artikel sumber pada studi kasus. Model yang dipakai adalah *checkpoint* RoBERTa berbahasa Indonesia yang tersedia publik [19] dan dilatih lanjut dengan resep yang sama pada setiap percobaan, sehingga yang diperiksa adalah efek *abstention* dan bukan perbandingan arsitektur. IndoNLI dipilih karena menyediakan dua tingkat kesulitan, yaitu *test_lay* dan *test_expert* [6].

Kontribusi penelitian ini ada tiga: (1) ukuran manfaat dan biaya *abstention* berbasis *threshold* pada NLI berbahasa Indonesia beserta variasinya antar *seed*; (2) pengujian apakah hasil itu berpindah ke klaim studi kasus beserta analisis kesalahan yang lolos; dan (3) pembanding *hypothesis-only* untuk menunjukkan batas pembacaan akurasi IndoNLI.

 ISSN:2302-8149
e-ISSN:2540-9719

2 Tinjauan Literatur

2.1 Verifikasi Klaim dan NLI

Guo dkk. [3] menyurvei pemeriksaan fakta otomatis dan menyimpulkan bahwa bidang ini masih menyisakan tantangan terbuka. Dataset FEVER [10] memuat 185.445 klaim yang dibuat dengan mengubah kalimat dari Wikipedia, masing-masing diberi putusan didukung, dibantah, atau informasi tidak cukup, yang sejajar dengan *entailment*, *contradiction*, dan *neutral* pada NLI. Schuster dkk. [4] membangun lebih dari 400 ribu pasangan klaim-bukti kontrastif dari revisi Wikipedia, dan pelatihan pada data itu menaikkan akurasi sebesar 10% pada verifikasi fakta adversarial dan 6% pada NLI adversarial. Honovich dkk. [5] membandingkan metode pendeteksi inkonsistensi faktual pada 11 dataset dan menemukan bahwa model NLI berskala besar termasuk yang berkinerja kuat. Karya-karya ini berbahasa Inggris dan, sejauh yang ditinjau, tidak menilai penolakan menjawab pada model NLI.

Label *neutral* tidak sama dengan *abstention*. Neutral adalah putusan tentang hubungan antara bukti dan klaim, sedangkan *abstention* adalah keputusan sistem tentang keyakinannya sendiri. Model dapat memprediksi *neutral* dengan yakin padahal salah, dan dapat ragu pada prediksi *entailment* yang benar. Karena itu penolakan menjawab perlu dinilai terpisah dari kemampuan model mengenali bukti yang tidak cukup.

2.2 NLI dan Deteksi Hoaks Berbahasa Indonesia

Untuk bahasa Indonesia, IndoNLI [6] memuat hampir 18 ribu pasangan yang ditulis anotator awam dan ahli. Bagian ahli hanya dipakai sebagai data uji dan menantang karena mencakup penalaran numerik, temporal, spasial, perubahan struktur kalimat, dan idiom. Model terbaik pada makalah itu, XLM-R, masih berselisih akurasi 13,4% dari manusia pada data uji ahli. Checkpoint RoBERTa berbahasa Indonesia yang dilatih pada IndoNLI tersedia publik dan melaporkan akurasi 77,06% pada validasi, 74,24% pada *test_lay*, dan 61,66% pada *test_expert* [19]. Di sisi hoaks, Fawaid dkk. [7] membandingkan CNN, BiLSTM, dan BERT pada 1.116 berita TurnBackHoax dengan akurasi terbaik 90%, Yefferson dkk. [8] dan Jocelynne dkk. [9] mencapai akurasi 93,2% dan 94,1% dengan IndoBERT, dan Arrahmi dkk. [20] memverifikasi klaim berbahasa Indonesia dengan penalaran LLM berbasis bukti (akurasi 0,92; F1 0,87).

Tabel 1 merangkum penelitian yang paling dekat dengan penelitian ini. Tidak satu pun dari penelitian itu menilai penolakan menjawab, sejauh yang ditinjau.

Tabel 1. Ringkasan penelitian terkait.

Penelitian	Data	Metode	Hasil terbaik	Menilai *abstention*
Fawaid dkk. [7]	1.116 berita TurnBackHoax	CNN, BiLSTM, CNN-BiLSTM, BERT	Akurasi 90% (BERT)	Tidak dilaporkan
Yefferson dkk. [8]	5.876 artikel TurnBackHoax dan CNN Indonesia	IndoBERT dan LSTM	Akurasi 93,2%	Tidak dilaporkan
Jocelynne dkk. [9]	1.302 artikel hoaks politik per kelas	*Fine-tuning* IndoBERT	Akurasi 94,1%	Tidak dilaporkan
Arrahmi dkk. [20]	Versi Indonesia dari X-FACT	Penalaran LLM berbasis bukti	Akurasi 0,92, F1 0,87	Tidak dilaporkan
Mahendra dkk. [6]	IndoNLI, hampir 18 ribu pasangan	XLM-R sebagai model terbaik	Selisih akurasi 13,4% dari manusia pada *test_expert*	Tidak dilaporkan
Penelitian ini	IndoNLI dan 758 klaim (270 hoaks, 488 valid)	RoBERTa NLI dengan *threshold*	Lihat Hasil dan Pembahasan	Ya

2.3 Pintasan pada Data NLI

Akurasi setinggi itu pada satu korpus belum menunjukkan bahwa model memeriksa isi klaim. Klasifikator yang hanya melihat hipotesis dapat melampaui tebakan kelas mayoritas pada banyak

 ISSN:2302-8149
e-ISSN:2540-9719

dataset NLI [21], dengan akurasi sekitar 67% pada SNLI dan 53% pada MultiNLI [22]. Gururangan dkk. mengaitkan hasil itu dengan pola penulisan anotator, misalnya kata negasi yang berkorelasi dengan label *contradiction*. Karena hipotesis IndoNLI juga ditulis oleh anotator [6], pola serupa mungkin ada, dan penelitian ini mengukurnya langsung dengan pembanding *hypothesis-only*. Pada data hoaks yang dikumpulkan sendiri, label berkorelasi penuh dengan sumber artikel (Subbab 3.2), sehingga data itu tidak dipakai untuk pelatihan.

2.4 Prediksi Selektif dan Kalibrasi

Gagasan menolak menjawab berasal dari opsi *reject* pada pengenalan pola [23]. Geifman dan El-Yaniv [24] membawanya ke jaringan saraf dalam melalui pengklasifikasi selektif yang tingkat risikonya dapat ditetapkan pengguna. Hendrycks dan Gimpel [14] menunjukkan bahwa probabilitas *softmax* maksimum menjadi pembanding dasar yang kuat untuk mendeteksi contoh yang salah diklasifikasi dan contoh di luar distribusi, dengan AUROC sebagai ukuran seberapa baik probabilitas itu memisahkan contoh benar dari salah. Guo dkk. [15] menunjukkan bahwa jaringan saraf modern kurang terkalibrasi dan bahwa *temperature scaling* yang sederhana memperbaikinya. Karena *threshold* pada probabilitas maksimum bergantung pada kalibrasi, penelitian ini memeriksa diskriminasi lewat AUROC dan kalibrasi lewat expected calibration error (ECE), tanpa memperbaiki kalibrasi.

Pada pemrosesan bahasa alami, Xin dkk. [12] membandingkan model dan penduga *confidence* untuk prediksi selektif dan melaporkan bahwa transformer terlatih meningkatkan akurasi sekaligus estimasi *confidence*. Varshney dkk. [16] menguji pendekatan prediksi selektif pada 17 dataset dalam kondisi dalam domain, luar domain, dan adversarial: tidak ada metode yang konsisten dan jauh mengungguli *baseline* probabilitas maksimum, dan kinerja tidak berpindah antar tugas. Jiang dkk. [25] menemukan bahwa T5, BART, dan GPT-2 tidak terkalibrasi dengan baik pada penjawab pertanyaan. Pada model bahasa besar, Wen dkk. [13] menyurvei penelitian *abstention* dan Feng dkk. [26] menaikkan akurasi menolak jawaban hingga 19,3% melalui kolaborasi antar model. Temuan-temuan ini menunjukkan bahwa probabilitas maksimum adalah titik awal yang wajar, tetapi efeknya harus diuji pada setiap tugas dan domain.

2.5 Variasi Antar Seed dan Celah Penelitian

Mosbach dkk. [17] menemukan bahwa hasil *fine-tuning* BERT, RoBERTa, dan ALBERT pada GLUE sangat bervariasi antar *seed*, dan Bouthillier dkk. [18] menunjukkan bahwa pengambilan data, inisialisasi parameter, dan pilihan hiperparameter sama-sama berpengaruh besar pada perbandingan model. Karena itu, hasil satu *seed* tidak cukup untuk menyimpulkan efek *abstention*.

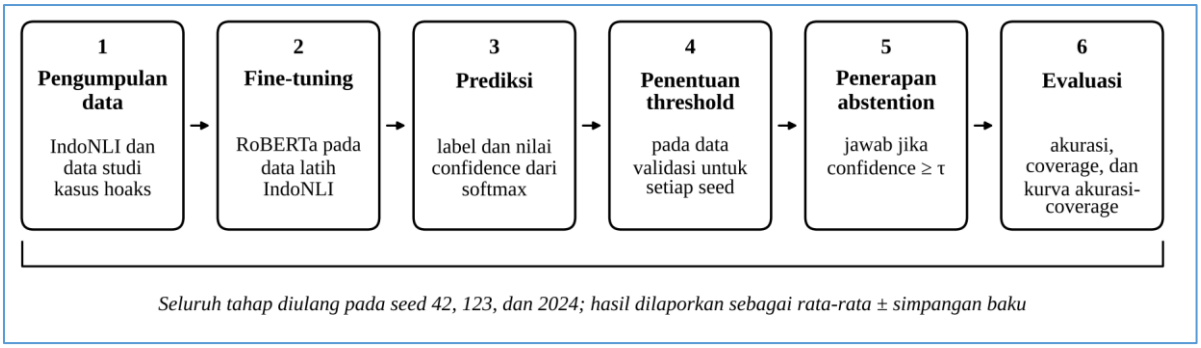
Dari tinjauan tersebut tampak dua celah. Pertama, penelitian hoaks berbahasa Indonesia yang ditinjau mengklasifikasikan teks berita atau memverifikasi klaim dengan LLM, dan tidak menilai penolakan menjawab. Kedua, bukti tentang prediksi selektif pada pemrosesan bahasa alami bergantung pada tugas dan domain, dan evaluasi untuk NLI berbahasa Indonesia tidak ditemukan. Penelitian ini mengisi kedua celah itu dengan mengevaluasi *abstention* berbasis *threshold* pada model NLI berbahasa Indonesia dengan tiga *seed*, pada IndoNLI dan data studi kasus.

3 Metode Penelitian

3.1 Alur Penelitian

Penelitian ini menguji apakah penolakan menjawab (*abstention*) berbasis ambang kepercayaan (*threshold*) meningkatkan keandalan model *natural language inference* (NLI) berbahasa Indonesia untuk verifikasi klaim hoaks. Klaim diperlakukan sebagai hipotesis dan bukti pendukung sebagai premis, dan model memprediksi *entailment* (bukti mendukung klaim), *contradiction* (bukti membantah klaim), atau *neutral* (bukti tidak cukup). Penelitian terdiri atas enam tahap seperti pada Gambar 1, dan seluruh tahap diulang pada tiga *seed* acak (42, 123, dan 2024) agar variasi akibat inisialisasi dan urutan data dapat diukur.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 1. Alur penelitian.

3.2 Data

Data utama adalah IndoNLI [6], korpus NLI berbahasa Indonesia berupa pasangan premis-hipotesis dengan label *entailment*, *neutral*, dan *contradiction*. Korpus ini terbagi menjadi data latih, data validasi, dan dua data uji, seperti dirinci pada Tabel 2. Data *test_lay* ditulis oleh anotator awam, sedangkan *test_expert* ditulis oleh anotator ahli dan dirancang lebih sulit. Data latih hanya dipakai untuk melatih model. Data validasi dipakai untuk memilih epoch terbaik dan menentukan *threshold*, sedangkan kedua data uji tidak pernah dipakai untuk keputusan apa pun.

Tabel 2. Pembagian data yang digunakan.

Data	Fungsi	Pasangan	Label
IndoNLI latih	Melatih model	10.330	3 kelas
IndoNLI validasi	Memilih epoch dan menentukan *threshold*	2.197	3 kelas
IndoNLI *test_lay*	Evaluasi, anotator awam	2.201	3 kelas
IndoNLI *test_expert*	Evaluasi, anotator ahli	2.984	3 kelas
Studi kasus (hoaks dan berita valid)	Evaluasi di domain studi kasus, tanpa pelatihan	758	*entailment* 488, *contradiction* 270

Selain IndoNLI, penelitian ini membangun data studi kasus untuk melihat perilaku model pada klaim di luar IndoNLI. Klaim berlabel *contradiction* dikumpulkan dari 400 artikel bantahan TurnBackHoax [27]: narasi hoaks menjadi hipotesis dan penjelasan pemeriksa fakta menjadi premis. Klaim berlabel *entailment* dikumpulkan dari 598 berita Antara [28]: judul berita menjadi hipotesis dan beberapa paragraf pembukanya menjadi premis. Dari 998 pasangan awal, 112 dibuang karena bocoran label atau kegagalan ekstraksi dan 128 karena duplikat, sehingga tersisa 758 pasangan, yaitu 270 hoaks (*contradiction*) dan 488 berita valid (*entailment*). Pengumpulan dilakukan pada Oktober 2026. Istilah klaim hoaks dalam artikel ini hanya merujuk pada 270 pasangan berlabel *contradiction*. Label diturunkan dari sumber artikel, bukan anotasi manual, dan data tidak memuat kelas *neutral*. Label berkorelasi penuh dengan sumber (seluruh *contradiction* dari TurnBackHoax, seluruh *entailment* dari Antara), sehingga ciri sumber tidak dapat dipisahkan dari hubungan semantik. Untuk mengukur besarnya, pengklasifikasi TF-IDF (unigram dan bigram) dengan regresi logistik dievaluasi dengan validasi silang lima lipatan pada 758 pasangan. Pengklasifikasi yang hanya menerima klaim mencapai akurasi 90,90 ± 1,10% dan yang hanya menerima bukti 98,02 ± 0,66%, dibanding tebakan mayoritas 64,38%. Kata pembeda terkuat pada bukti adalah penanda sumber, misalnya “antara” dan “jakarta” untuk *entailment* serta “tim pemeriksa fakta” dan “google” untuk *contradiction*, sehingga pembersihan teks mengurangi tetapi tidak menghilangkan ciri sumber. Karena itu data studi kasus tidak dipakai untuk pelatihan atau penyetelan: model dilatih hanya pada IndoNLI dan diterapkan tanpa pelatihan ulang, sehingga tidak mempelajari ciri sumber dari label ini. Namun ciri sumber tetap dapat memengaruhi prediksi secara tidak langsung, sehingga akurasi pada data ini tidak dibaca sebagai kemampuan menalar isi klaim, melainkan sebagai perilaku model dan *abstention* pada teks dari domain yang berbeda.

 ISSN:2302-8149
e-ISSN:2540-9719

3.3 Model dan Pelatihan

Model yang digunakan adalah w11wo/indonesian-roberta-base-indonli [19], model RoBERTa [29] berukuran base dengan sekitar 125 juta parameter yang menghasilkan tiga kelas. Premis dan hipotesis digabung menjadi satu masukan dan dipotong sampai 256 token. Urutan label pada *checkpoint* (*neutral*, *contradiction*, *entailment*) berbeda dari IndoNLI (*entailment*, *neutral*, *contradiction*). Penelitian ini memakai urutan IndoNLI pada seluruh proses: label data dipakai dalam indeks asli IndoNLI, konfigurasi model diatur ke urutan itu, dan prediksi diterjemahkan kembali dengan urutan yang sama. Bobot lapisan klasifikasi tidak ditukar, sehingga pada awal pelatihan indeks keluarannya tergeser secara siklik terhadap label data, dan pemetaan yang benar dipelajari ulang selama pelatihan lanjutan karena seluruh bobot ikut diperbarui. Kebenaran pemetaan diverifikasi dua kali. Pertama, *checkpoint* asli tanpa pelatihan lanjutan mencapai akurasi validasi hanya 12,34% (di bawah tebakan acak 33%) jika indeks keluarannya dibaca langsung sebagai indeks IndoNLI, tetapi 76,70% jika dipetakan ke urutan IndoNLI. Kedua, model hasil pelatihan lanjutan mencapai akurasi validasi 77,92%, dengan prediksi benar terbanyak pada diagonal matriks konfusi untuk ketiga kelas.

Checkpoint ini sudah dilatih pada IndoNLI oleh pengembangnya selama lima epoch [19], sehingga pelatihan di sini bersifat lanjutan. Kartu model melaporkan akurasi 77,06% (dev), 74,24% (*test_lay*), dan 61,66% (*test_expert*), tetapi tidak menyebut split pelatihan dan skripnya tidak dipublikasikan, sehingga ketiadaan kebocoran tidak dapat dipastikan dari dokumentasi. Penelitian ini memakai split resmi IndoNLI apa adanya dan memeriksa kebocoran secara empiris. Checkpoint asli tanpa pelatihan lanjutan mencapai akurasi 88,10% pada 2.000 sampel data latih, tetapi hanya 76,70% pada validasi, 74,01% pada *test_lay*, dan 58,38% pada *test_expert*. Akurasi pada tiga split terakhir sedikit di bawah angka kartu model dan jauh di bawah akurasi data latih, sehingga tidak ada tanda bahwa split tersebut dihafal. Pemeriksaan ini tidak membuktikan ketiadaan kebocoran, dan risiko yang tersisa dibahas pada Keterbatasan. Data studi kasus tidak berasal dari IndoNLI sehingga tidak termasuk data latih *checkpoint*.

Model dilatih lebih lanjut pada data latih IndoNLI selama 3 epoch dengan ukuran *batch* 16 dan laju pembelajaran 2×10^{-5} (Tabel 3). Epoch yang dipakai adalah epoch dengan F1 makro tertinggi pada data validasi. Pelatihan dijalankan di Google Colab dengan GPU menggunakan pustaka *transformers* 5.18.0 dan *datasets* 4.8.5 [30].

Tabel 3. Konfigurasi pelatihan.

Parameter	Nilai
Model dasar	*w11wo/indonesian-roberta-base-indonli*
Jumlah kelas	3 (*entailment*, *neutral*, *contradiction*)
Panjang masukan maksimum	256 token
Jumlah epoch	3
Ukuran *batch*	16
Laju pembelajaran	2×10^{-5}
Pemilihan model	F1 makro validasi tertinggi
Seed	42, 123, 2024
Pengoptimal	AdamW (bawaan *Trainer*)
Weight decay dan *warmup*	0 dan tanpa *warmup*
Penjadwal laju pembelajaran	Linear

Hasil pelatihan bergantung pada faktor acak, yaitu urutan data dan *dropout* [17], [18]. Karena itu *seed* dikunci pada dua tempat: sebelum model dimuat, dan pada argumen pelatihan untuk inisialisasi serta pengacakan data. Pelatihan dan seluruh evaluasi diulang untuk *seed* 42, 123, dan 2024.

3.4 Abstention Berbasis Threshold

Untuk satu pasangan premis-hipotesis x, model menghasilkan skor mentah z untuk setiap kelas. Skor ini diubah menjadi probabilitas dengan fungsi *softmax* pada Persamaan (1), dan nilai *confidence* didefinisikan sebagai probabilitas terbesar pada Persamaan (2).

 ISSN:2302-8149
e-ISSN:2540-9719

$$p_k(x) = \frac{\exp(z_k)}{\sum_{j=1}^{3} \exp(z_j)} \quad (1)$$

$$c(x) = \max_k p_k(x) \quad (2)$$

Model hanya menjawab jika *confidence* mencapai *threshold* τ, dan menolak menjawab jika tidak, seperti pada Persamaan (3), dengan $\hat{y}(x)$ adalah kelas berprobabilitas terbesar. Karena tiga kelas, nilai *confidence* paling kecil adalah 1/3, sehingga $\tau \leq 1/3$ sama dengan model tanpa *abstention* dan menjadi pembanding dalam penelitian ini.

$$a(x) = \begin{cases} \hat{y}(x), & c(x) \geq \tau \\ \text{abstain}, & c(x) < \tau \end{cases} \quad (3)$$

Nilai *confidence* yang sama tidak berarti sama pada model yang berbeda, karena kalibrasi bergantung pada *seed*. Sebagai contoh, τ = 0,8 menghasilkan *coverage* validasi 93,76% pada *seed* 42 tetapi hanya 77,74% pada *seed* 123, sehingga τ tidak ditetapkan sebagai angka yang sama untuk semua *seed*. Penelitian ini menetapkan *coverage* sasaran pada data validasi, yaitu 90% sebagai pengaturan utama serta 95% dan 80% sebagai analisis sensitivitas, lalu mengambil τ sebagai kuantil ke-(1 − sasaran) dari *confidence* data validasi pada setiap *seed*. Nilai τ yang diperoleh diterapkan tanpa perubahan pada data uji dan data studi kasus.

3.5 Skenario Evaluasi dan Metrik

Evaluasi dilakukan pada data validasi, *test_lay*, *test_expert*, dan studi kasus. Tanpa *abstention*, model dinilai dengan akurasi, sedangkan F1 makro dipakai untuk memilih epoch. Dengan *abstention*, model dinilai dengan dua ukuran: *coverage*, yaitu proporsi pasangan yang dijawab, pada Persamaan (4), dan akurasi pada pasangan yang dijawab, pada Persamaan (5). Pada kedua persamaan, *A* adalah himpunan pasangan yang dijawab, *N* adalah jumlah seluruh pasangan, dan *y*(*x*) adalah label sebenarnya.

$$\text{coverage} = \frac{|A|}{N}, \quad A = \{x : c(x) \geq \tau\} \quad (4)$$

$$\text{acc}_{\text{answered}} = \frac{1}{|A|} \sum_{x \in A} 1\big(\hat{y}(x) = y(x)\big) \quad (5)$$

Manfaat *abstention* diukur dengan selisih akurasi pada pasangan yang dijawab terhadap akurasi tanpa *abstention* pada *seed* yang sama, dihitung per *seed* lalu dilaporkan sebagai rata-rata ± simpangan baku sampel (n − 1) dari tiga *seed*. Kurva akurasi terhadap *coverage* dibuat dengan mengurutkan prediksi dari *confidence* tertinggi, lalu menghitung akurasi pada k prediksi teratas dengan k mencakup 50% sampai 100% data. Pada studi kasus, akurasi juga dilaporkan per kelas, ditambah porsi kesalahan yang berupa prediksi *neutral*, karena data ini tidak memuat kelas *neutral*.

Untuk menilai apakah *confidence* memisahkan prediksi benar dari salah, dihitung AUROC dengan prediksi benar sebagai kelas positif dan *confidence* sebagai skor, yaitu peluang sebuah prediksi benar yang dipilih acak memiliki *confidence* lebih tinggi daripada prediksi salah yang dipilih acak (0,5 berarti tidak berguna, 1 berarti pemisahan sempurna). AUROC mengukur diskriminasi dan tidak mengukur kalibrasi: nilainya tidak berubah jika seluruh *confidence* ditransformasi secara monoton, sehingga model dengan AUROC tinggi masih dapat terlalu yakin. Kalibrasi diukur dengan *expected calibration error* (ECE) [15], yang dihitung pada Persamaan (6).

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (6)$$

Pada Persamaan (6), *M* adalah jumlah *bin confidence* dengan lebar sama (15 pada penelitian ini), B_m adalah himpunan pasangan dalam *bin* ke-*m*, *acc*(B_m) adalah akurasi, *conf*(B_m) adalah rata-rata *confidence* dalam *bin* tersebut, dan *N* adalah jumlah seluruh pasangan. Pada studi kasus, prediksi *neutral* dihitung salah. Analisis kesalahan pada studi kasus membandingkan pasangan yang dijawab dan yang ditolak pada *threshold* sasaran 90%, memeriksa komposisi kesalahan yang lolos, serta membandingkan akurasi dengan porsi kelas mayoritas. Persentase pada analisis ini dibedakan menurut penyebutnya. Hoaks ditolak adalah persentase dari seluruh 270 hoaks. Hoaks salah prediksi adalah hoaks yang diprediksi *entailment* atau *neutral*, dan persentase yang lolos dihitung terhadap seluruh hoaks salah prediksi pada *seed* tersebut. Hoaks yang dinilai didukung bukti adalah bagian dari hoaks salah prediksi yang diprediksi *entailment*, dan persentase yang ditolak dihitung terhadap bagian

 ISSN:2302-8149
e-ISSN:2540-9719

itu. Kesalahan yang lolos adalah semua prediksi salah pada pasangan yang dijawab, baik hoaks maupun klaim valid. Untuk pembacaan kualitatif, pada setiap *seed* diambil 15 kesalahan prediksi *neutral* yang lolos dengan *confidence* tertinggi dan 15 pasangan yang ditolak dengan *confidence* terendah, sehingga terkumpul 90 baris yang dibaca manual dan tidak mewakili seluruh data.

Sebagai pembanding, *baseline hypothesis-only* [21], [22] melatih model dengan resep yang sama tetapi hanya menerima hipotesis, dengan panjang masukan maksimum 128 token. *Baseline* ini dijalankan pada tiga *seed* (42, 123, dan 2024), dan dibandingkan dengan tebakan kelas mayoritas serta model utama. Model yang sama diterapkan pada klaim di studi kasus tanpa dilatih ulang, untuk melihat apakah pola hipotesis dari IndoNLI berlaku pada klaim hoaks.

4 Hasil dan Pembahasan

Hasil disajikan mengikuti urutan langkah pada Gambar 1. Seluruh angka adalah rata-rata ± simpangan baku sampel dari tiga *seed* (42, 123, dan 2024), kecuali dinyatakan lain.

4.1 Data

Pembanding *hypothesis-only* menunjukkan bahwa label IndoNLI dapat ditebak dari hipotesis saja dengan akurasi di atas tebakan mayoritas (Tabel 4). Rata-rata tiga *seed*, model yang hanya melihat hipotesis mencapai akurasi 59,25 ± 0,31% pada data validasi dan 59,26 ± 0,29% pada *test_lay*, atau 22,52 dan 22,55 poin di atas tebakan kelas mayoritas (36,73% dan 36,71%). Pada *test_expert* akurasinya 46,73 ± 0,35%, 11,84 poin di atas tebakan mayoritas (34,89%). Model utama mencapai 77,47%, 73,39%, dan 60,27%, sehingga pembanding ini mencakup sekitar 55% dari keunggulan model utama atas tebakan mayoritas pada data validasi, 61% pada *test_lay*, dan 47% pada *test_expert*. Simpangan baku antar *seed* hanya 0,29 sampai 0,35 poin.

Tabel 4. Hasil pembanding hypothesis-only (akurasi dan F1 makro, %, rata-rata ± simpangan baku tiga seed).

Data	Tebakan mayoritas	*Hypothesis-only* akurasi	*Hypothesis-only* F1 makro	Model utama
Validasi	36,73	59,25 ± 0,31	59,10 ± 0,46	77,47 ± 0,39
test_lay	36,71	59,26 ± 0,29	59,02 ± 0,43	73,39 ± 0,43
test_expert	34,89	46,73 ± 0,35	46,67 ± 0,33	60,27 ± 1,00
Studi kasus	64,38	57,96 ± 1,61	30,15 ± 1,04	62,14 ± 9,78

Pola ini sejalan dengan temuan pada SNLI dan MultiNLI, tempat klasifikator yang hanya melihat hipotesis mencapai sekitar 67% dan 53% [22]. Hasil ini menunjukkan bahwa akurasi IndoNLI tidak seluruhnya berasal dari penalaran atas premis, dan porsi pola hipotesis paling kecil pada *test_expert* yang ditulis anotator ahli (47%, dibanding 55% dan 61% pada dua data lain). Angkanya berasal dari model yang dimulai dari *checkpoint* yang sudah pernah dilatih pada pasangan lengkap IndoNLI, sehingga dibaca sebagai perkiraan kasar.

Pada klaim di studi kasus, model *hypothesis-only* memprediksi *entailment* untuk 81,9 ± 2,1% klaim hoaks dan 83,8 ± 2,1% klaim valid. Akurasinya 57,96 ± 1,61% dan F1 makronya 0,302 ± 0,010, di bawah tebakan mayoritas (64,38%) pada ketiga *seed*. Hasil ini hanya menunjukkan bahwa pola hipotesis yang dipelajari dari IndoNLI tidak berpindah ke klaim studi kasus, karena model tidak dilatih pada data tersebut.

4.2 Model dan Pelatihan

Tanpa *abstention*, akurasi model stabil antar *seed* pada data IndoNLI tetapi tidak pada studi kasus (Tabel 5). Simpangan baku antar *seed* hanya 0,39 poin pada validasi, 0,43 poin pada *test_lay*, dan 1,00 poin pada *test_expert*, dan akurasinya berselisih kecil dari kartu model (77,06%, 74,24%, dan 61,66% [19]). Pada studi kasus, akurasi berkisar dari 55,41% (*seed* 123) sampai 73,35% (*seed* 2024), dengan rata-rata 62,14 ± 9,78%, yaitu simpangan baku 9,8 sampai 25,1 kali lebih besar daripada pada data IndoNLI.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 5. Akurasi tanpa abstention untuk setiap seed (%).

Seed	Validasi	*test_lay*	*test_expert*	Studi kasus
42	77,24	73,83	60,46	57,65
123	77,24	72,97	59,18	55,41
2024	77,92	73,38	61,16	73,35
Rata-rata ± std	77,47 ± 0,39	73,39 ± 0,43	60,27 ± 1,00	62,14 ± 9,78

Epoch terbaik juga berbeda antar *seed*. Model *seed* 123 berasal dari epoch 1 (F1 makro validasi 0,7704, *validation loss* 0,67), sedangkan model *seed* 42 berasal dari epoch 3 (F1 makro 0,7713, *validation loss* 1,11). Akurasi validasi keduanya sama (77,24%), tetapi tingkat *confidence* model berbeda, dan perbedaan ini berdampak pada *threshold* seperti dibahas berikut.

4.3 Abstention Berbasis Threshold

Pemetaan dari *threshold* ke *coverage* berbeda antar *seed*, sehingga τ diambil dari *coverage* sasaran pada data validasi. Untuk sasaran 90%, τ bernilai 0,904 pada *seed* 42, 0,625 pada *seed* 123, dan 0,937 pada *seed* 2024. Perbedaan ini mengikuti sebaran *confidence* (Tabel 6): contoh validasi dengan *confidence* di atas 0,99 berjumlah 74,6% pada *seed* 42 dan 77,2% pada *seed* 2024, tetapi hanya 0,5% pada *seed* 123, dengan median *confidence* prediksi salah 0,988; 0,992; dan 0,823.

Tabel 6. Sebaran confidence pada data validasi dan threshold sasaran 90%.

Ukuran	*Seed* 42	*Seed* 123	*Seed* 2024
τ untuk sasaran *coverage* 90%	0,904	0,625	0,937
Contoh dengan *confidence* > 0,99 (%)	74,6	0,5	77,2
Contoh dengan *confidence* > 0,999 (%)	0,3	0,0	26,9
Median *confidence* prediksi benar	0,998	0,962	0,996
Median *confidence* prediksi salah	0,988	0,823	0,992

Seed 2024 memiliki 26,9% contoh validasi dengan *confidence* di atas 0,999, sehingga kurvanya mendatar di bawah *coverage* 70% (Gambar 2): akurasi pada 50% prediksi paling yakin hanya 83,99%, dibanding 91,63% pada *seed* 42 dan 90,99% pada *seed* 123. Penjelasan ini belum diuji secara langsung.

4.4 Skenario Evaluasi dan Metrik

Pada sasaran *coverage* 90%, akurasi pada data yang dijawab naik di keempat data (Tabel 7), dengan *coverage* dan akurasi dihitung memakai Persamaan (4) dan (5). Pada data IndoNLI, kenaikannya 3,12 ± 0,27 poin pada validasi, 4,32 ± 0,48 poin pada *test_lay*, dan 3,60 ± 0,28 poin pada *test_expert*. Coverage nyata menurun mengikuti kesulitan data: 89,99% pada validasi, 87,60% pada *test_lay*, 84,41% pada *test_expert*, dan 71,46% pada studi kasus. Penyebabnya, τ ditetapkan pada validasi, sedangkan data yang lebih sulit atau asing memiliki lebih banyak contoh dengan *confidence* di bawah τ.

Tabel 7. Hasil abstention pada sasaran coverage validasi 90%.

Data	Tanpa *abstention* (%)	*Coverage* nyata (%)	Akurasi dijawab (%)	Selisih (poin)
Validasi	77,47 ± 0,39	89,99	80,59 ± 0,39	+3,12 ± 0,27
test_lay	73,39 ± 0,43	87,60 ± 0,60	77,71 ± 0,13	+4,32 ± 0,48
test_expert	60,27 ± 1,00	84,41 ± 1,79	63,87 ± 0,94	+3,60 ± 0,28
Studi kasus	62,14 ± 9,78	71,46 ± 2,15	69,50 ± 11,75	+7,36 ± 1,98

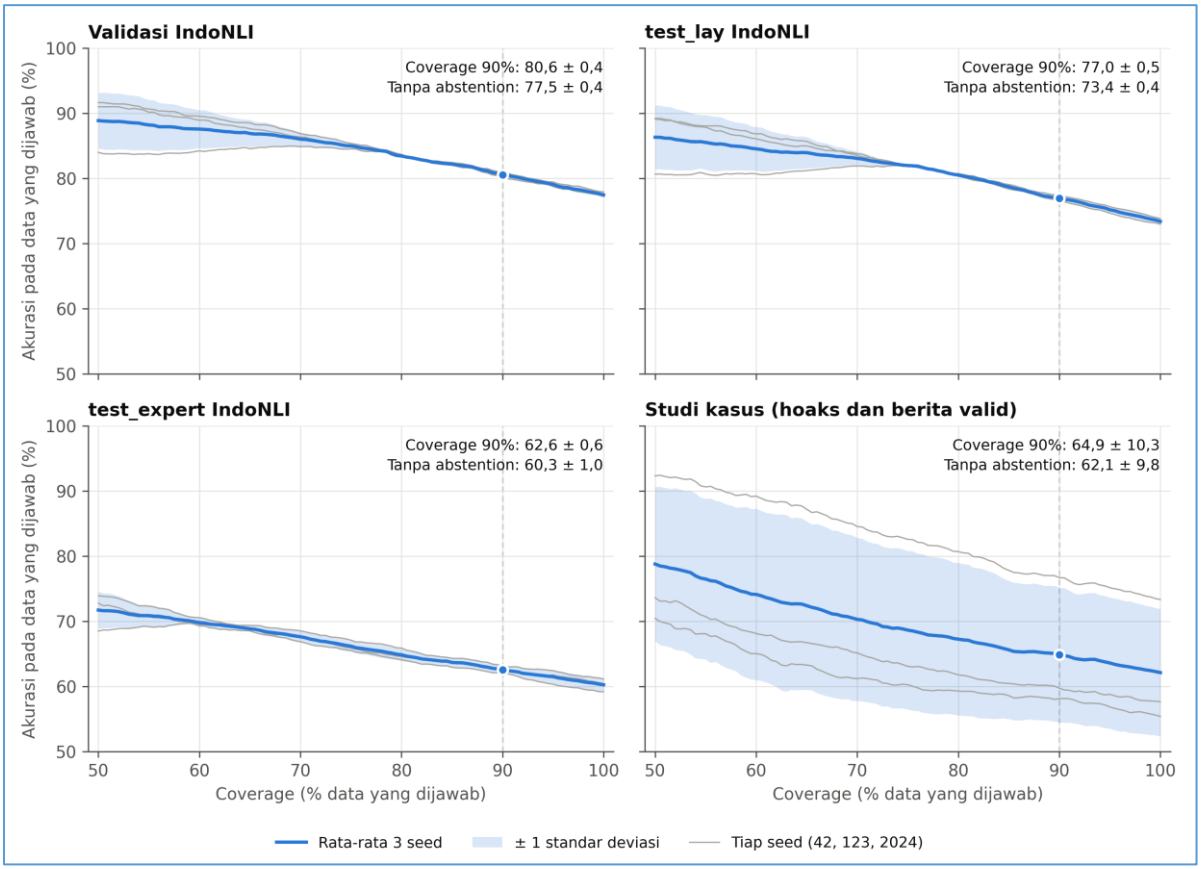
Menurunkan *coverage* sasaran menaikkan akurasi lebih besar (Tabel 8). Pada sasaran 95%, kenaikannya 1,59 sampai 3,90 poin. Pada sasaran 80%, kenaikannya 5,99 sampai 16,93 poin, tetapi *coverage* nyata turun menjadi 71,67% pada *test_expert* dan 50,31% pada studi kasus, sehingga kenaikan pada studi kasus harus dibaca bersama *coverage*-nya.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 8. Selisih akurasi pada sasaran coverage 95% dan 80%.

Data	*Coverage* nyata, sasaran 95% (%)	Selisih (poin)	*Coverage* nyata, sasaran 80% (%)	Selisih (poin)
Validasi	94,99	+1,59 ± 0,16	79,97	+5,99 ± 0,49
test_lay	93,29 ± 0,70	+2,44 ± 0,21	77,63 ± 0,79	+7,86 ± 0,46
test_expert	92,20 ± 0,84	+1,76 ± 0,18	71,67 ± 2,60	+6,78 ± 1,06
Studi kasus	83,69 ± 0,86	+3,90 ± 0,81	50,31 ± 3,76	+16,93 ± 1,14

Perbandingan pada *coverage* yang sama memberi gambaran lebih seimbang (Gambar 2). Pada *coverage* 90% yang dihitung pada masing-masing data, akurasi naik 3,10; 3,57; 2,29; dan 2,72 poin pada validasi, *test_lay*, *test_expert*, dan studi kasus. Pada *coverage* 70%, kenaikannya 8,55; 9,67; 7,37; dan 8,17 poin. Jadi kenaikan 7,36 poin pada studi kasus (Tabel 7) terutama mencerminkan *coverage* yang turun menjadi 71,5%, bukan seleksi yang lebih baik daripada data dalam domain.



Gambar 2. Kurva akurasi terhadap coverage pada empat data (rata-rata tiga seed dan ± 1 simpangan baku).

Confidence memisahkan prediksi benar dari salah lebih baik daripada acak (AUROC 0,5) tetapi belum sempurna (AUROC 1) (Tabel 9). AUROC rata-rata 0,732 pada validasi, 0,735 pada *test_lay*, dan 0,665 pada *test_expert*. Pada studi kasus, AUROC 0,734 ± 0,055, dalam kisaran data IndoNLI. Jadi urutan *confidence* berpindah ke data studi kasus, tetapi *coverage* tidak, yaitu *coverage* nyata hanya 71,5% dibanding sasaran 90%.

AUROC hanya mengukur diskriminasi dan tidak menunjukkan apakah nilai *confidence* sesuai dengan peluang benar. ECE menunjukkan bahwa model terlalu yakin pada semua data: rata-rata *confidence* 93,79% dengan akurasi 77,47% pada validasi, 92,91% dengan 73,39% pada *test_lay*, 91,40% dengan 60,27% pada *test_expert*, dan 85,44% dengan 62,14% pada studi kasus. ECE rata-rata 16,4%; 19,6%; 31,2%; dan 23,4% pada keempat data (Tabel 9). Jadi urutan *confidence* cukup baik, tetapi nilainya tidak dapat dibaca sebagai probabilitas benar, dan inilah alasan *threshold* ditetapkan dari *coverage* sasaran.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 9. AUROC dan ECE confidence pada setiap data dan seed.

Data	*Seed* 42	*Seed* 123	*Seed* 2024	Rata-rata ± std
AUROC (semakin tinggi semakin baik)				
Validasi	0,766	0,759	0,670	0,732 ± 0,053
test_lay	0,761	0,763	0,680	0,735 ± 0,047
test_expert	0,675	0,686	0,633	0,665 ± 0,028
Studi kasus	0,707	0,697	0,797	0,734 ± 0,055
ECE, dalam % (semakin rendah semakin baik)				
Validasi	19,4	10,7	19,2	16,4 ± 5,0
test_lay	22,2	13,4	23,1	19,6 ± 5,3
test_expert	34,9	24,6	34,0	31,2 ± 5,7
Studi kasus	33,1	18,1	19,0	23,4 ± 8,4

4.5 Studi Kasus Klaim Hoaks dan Berita Valid

Perilaku model pada klaim hoaks berbeda antar *seed*, dengan akurasi tanpa *abstention* berselisih 17,94 poin antara *seed* terendah dan tertinggi. Prediksi *neutral* mencakup 35,49% pasangan pada *seed* 42, 36,94% pada *seed* 123, dan hanya 15,96% pada *seed* 2024. Akurasi per kelas pada ketiga *seed* berturut-turut 36,30%, 35,56%, dan 61,48% untuk hoaks, serta 69,47%, 66,39%, dan 79,92% untuk klaim valid. Lonjakan akurasi *seed* 2024 berasal dari turunnya prediksi *neutral*, dan ikut menaikkan hoaks yang dinilai didukung bukti dari 8,52% dan 9,26% menjadi 18,89%. Akurasi rata-rata tanpa *abstention* (62,14%) berada di bawah tebakan kelas mayoritas, yaitu menjawab semua pasangan sebagai *entailment* (64,38%), dan hanya *seed* 2024 (73,35%) yang melampauinya.

Dengan τ dari validasi dan sasaran 90%, *abstention* menolak 27% sampai 31% pasangan studi kasus (Tabel 10). Akurasi pada pasangan yang ditolak (41,1%; 41,7%; dan 47,3%) lebih rendah 19,9 sampai 35,7 poin daripada yang dijawab (63,9%; 61,6%; dan 83,0%), sehingga seleksi bekerja. Hoaks lebih sering ditolak (37,4% sampai 48,1% dari 270) daripada klaim valid (20,3% sampai 22,1% dari 488). Kesalahan yang paling berbahaya adalah hoaks yang dinilai didukung bukti, yaitu hoaks yang diprediksi *entailment*. Dari 23, 25, dan 51 hoaks semacam itu pada ketiga *seed*, 8, 9, dan 26 lolos dan 15, 16, dan 25 ditolak, sehingga *abstention* menolak 65,2%; 64,0%; dan 49,0% di antaranya (Tabel 10), lebih tinggi daripada laju penolakan keseluruhan (27,6%; 31,0%; dan 27,0%).

Tabel 10. Pasangan studi kasus yang dijawab dan ditolak pada sasaran coverage 90%.

Ukuran	*Seed* 42	*Seed* 123	*Seed* 2024
τ	0,904	0,625	0,937
Coverage (%)	72,43	69,00	72,96
Akurasi dijawab (%)	63,9	61,6	83,0
Akurasi ditolak (%)	41,1	41,7	47,3
Hoaks ditolak (% dari 270)	37,4	48,1	39,3
Hoaks diprediksi entailment yang ditolak (%)	65,2 (15 / 23)	64,0 (16 / 25)	49,0 (25 / 51)
Klaim valid ditolak (% dari 488)	22,1	21,5	20,3
Kesalahan yang lolos / dijawab	198 / 549	201 / 523	94 / 553
Kesalahan lolos berupa *neutral* (%)	88,9	89,6	58,5
Hoaks salah prediksi yang lolos (%)	65,1 (112 / 172)	56,3 (98 / 174)	39,4 (41 / 104)

Namun 58,5% sampai 89,6% dari kesalahan yang lolos, pada hoaks maupun klaim valid, adalah prediksi *neutral* (176 dari 198 pada *seed* 42, 180 dari 201 pada *seed* 123, dan 55 dari 94 pada *seed* 2024). Di antara hoaks yang salah prediksi (172, 174, dan 104 pasangan), 65,1%, 56,3%, dan 39,4% tetap lolos (112, 98, dan 41 pasangan). Persentase ini berbeda dari angka penolakan 65,2%; 64,0%; dan 49,0% di atas karena penyebutnya hanya hoaks yang diprediksi *entailment*. Pembacaan manual atas 90 contoh menunjukkan dua pola pada kesalahan *neutral* yang lolos. Pertama, pada klaim Antara (34 dari 45 contoh), judul yang dijadikan klaim sering bukan pernyataan yang dapat diperiksa,

 ISSN:2302-8149
e-ISSN:2540-9719

sehingga prediksi *neutral* dengan *confidence* 0,93 sampai 0,999 dapat dibela dan label dari sumber menjadi penyebab kesalahan. Kedua, pada klaim TurnBackHoax (11 dari 45 contoh), tujuh dari delapan klaim yang berbeda berupa penggalan kalimat, dan premisnya tidak menyatakan klaim salah dalam 400 karakter pertama yang tersimpan, sedangkan model menerima sampai 256 token. Dari 24 klaim berbeda pada kelompok ini, 14 muncul pada lebih dari satu *seed*, sehingga kesalahan ini sistematis.

Pada 45 pasangan yang ditolak dengan *confidence* terendah (0,34 sampai 0,51), 15 pasangan (33%) sebenarnya diprediksi benar, sehingga *abstention* menahan sebagian jawaban yang benar.

4.6 Pembahasan

Hasil pada data dalam domain mendukung penggunaan probabilitas *softmax* maksimum sebagai dasar *abstention*, sejalan dengan temuan bahwa pembanding ini kuat [14]. Kenaikan akurasi konsisten antar *seed* dan membesar ketika *coverage* diturunkan, tetapi bentuknya pertukaran: pada sasaran 90%, 10,01% pasangan validasi diserahkan ke pemeriksa manusia untuk menaikkan akurasi 3,12 poin. Penetapan τ dari *coverage* sasaran diperlukan karena τ untuk sasaran yang sama berkisar dari 0,625 sampai 0,937 antar *seed*, konsisten dengan temuan bahwa jaringan saraf tidak selalu terkalibrasi [15] dan hasil *fine-tuning* bervariasi antar *seed* [17], [18]. Kalibrasi ulang seperti temperature scaling [15] tidak diuji.

Pada data studi kasus, urutan *confidence* berpindah tetapi *coverage* tidak. AUROC 0,73 setara dengan data dalam domain, dan kenaikan akurasi pada *coverage* yang sama sebanding. Hal ini tidak bertentangan dengan temuan bahwa keunggulan metode prediksi selektif tidak berpindah antar domain [16], karena yang diuji hanya satu metode pada satu pasang domain. Namun τ dari validasi menghasilkan *coverage* hanya 71,5%, dan akurasi pada pasangan yang dijawab (69,50%) tidak melampaui porsi kelas *entailment* pada pasangan yang sama, yaitu 70,9% rata-rata (69,2%, 73,2%, dan 70,3%). Hanya *seed* 2024 (83,0%) yang melampauinya.

Variasi antar *seed* pada studi kasus (55,41% sampai 73,35%, selisih 17,94 poin) jauh lebih besar daripada pada data dalam domain, padahal seluruh *seed* dievaluasi pada 758 pasangan yang sama, sehingga perbedaan berasal dari model dan bukan dari sampel data. Data yang ada menunjuk pada cara model memutuskan pada teks di luar domain. Prediksi *neutral* hanya 15,96% pada *seed* 2024, dibanding 35,49% dan 36,94% pada dua *seed* lain. Karena studi kasus tidak memuat label *neutral*, setiap prediksi *neutral* dihitung salah, sehingga *seed* yang jarang memilih *neutral* memperoleh akurasi lebih tinggi: 79,92% pada klaim valid dan 61,48% pada hoaks untuk *seed* 2024, dibanding 66,39% dan 35,56% untuk *seed* 123. Penyebab pastinya belum diuji, misalnya apakah berasal dari inisialisasi lapisan klasifikasi, urutan data, atau epoch yang terpilih. Kenaikan akurasi pada *seed* 2024 juga bukan pemeriksaan klaim yang lebih baik, karena hoaks yang dinilai didukung bukti naik dari 8,52% dan 9,26% menjadi 18,89%.

Variasi ini berdampak pada keandalan model dalam tiga hal. Pertama, akurasi dalam domain yang hampir sama (77,24% sampai 77,92% pada validasi) tidak memprediksi perilaku di luar domain. Kedua, kesimpulan dari satu *seed* dapat berlawanan: *seed* 2024 melampaui tebakan kelas mayoritas (73,35% dibanding 64,38%) dan porsi *entailment* pada pasangan yang dijawab (83,0% dibanding 70,3%), sedangkan *seed* 42 dan 123 tidak. Ketiga, *threshold* dari validasi tidak menjamin hasil di luar domain: pada *coverage* nyata 69,0% sampai 73,0%, akurasi pada pasangan yang dijawab berkisar 61,6% sampai 83,0%. Karena itu, penggunaan pada klaim hoaks memerlukan evaluasi pada beberapa *seed* dengan rentang hasil yang dilaporkan, dan data validasi dari domain sasaran yang dianotasi manual untuk menetapkan *threshold*. Penggabungan beberapa model dan kalibrasi ulang mungkin mengurangi variasi ini, tetapi belum diuji.

Prediksi *neutral* mencakup 58,5% sampai 89,6% kesalahan yang lolos. Prediksi ini tidak menyesatkan seperti hoaks yang dinilai benar, dan pada klaim yang tidak dapat diperiksa atau terpotong, *neutral* adalah jawaban yang masuk akal. Karena data tidak memuat kelas *neutral*, akurasi studi kasus kemungkinan menilai kinerja model terlalu rendah, dan label yang diturunkan dari sumber artikel menjadi batas utama evaluasi. Sebaliknya, 39% sampai 65% hoaks yang salah prediksi tetap lolos, sehingga *abstention* berbasis *confidence* hanya membantu menyaring dan tidak menggantikan pemeriksa manusia pada klaim hoaks.

Pembanding *hypothesis-only* menunjukkan bahwa 47% sampai 61% keunggulan model utama atas tebakan mayoritas dapat dicapai tanpa premis [21], [22], sehingga akurasi 77% pada validasi

tidak dapat dibaca sebagai kemampuan memeriksa klaim terhadap bukti sepenuhnya. Pada studi kasus, pola hipotesis dari IndoNLI tidak berpindah, tetapi ini belum membuktikan bahwa label studi kasus bebas dari pintasan sumber, karena pengklasifikasi TF-IDF sederhana mencapai akurasi 90,90% dari klaim saja dan 98,02% dari bukti saja (Subbab 3.2). Akurasi 90% sampai 94,1% pada penelitian deteksi hoaks sebelumnya [7], [8], [9] tidak dapat dibandingkan langsung dengan 62,14% di sini karena tugasnya klasifikasi artikel, bukan pemeriksaan klaim terhadap bukti.

Keterbatasan penelitian ini adalah sebagai berikut. Pertama, label studi kasus diturunkan dari sumber artikel tanpa anotasi manual dan tanpa kelas *neutral*, serta berkorelasi penuh dengan sumber itu sehingga ciri sumber tidak terpisah dari hubungan semantik. Karena itu temuan studi kasus hanya berlaku untuk pemisahan klaim dan bukti dari dua sumber ini dan tidak digeneralisasi ke kemampuan verifikasi hoaks secara umum. Ketiadaan label *neutral* juga membuat setiap prediksi *neutral* dihitung salah, padahal sebagian klaim, misalnya yang terpotong akibat ekstraksi, memang tidak dapat diperiksa dari bukti yang tersedia, sehingga akurasi studi kasus dapat lebih rendah daripada kemampuan model yang sebenarnya. Kedua, data dikumpulkan sekali, hanya tiga *seed* dipakai sehingga simpangan baku sampel kurang stabil, dan 90 contoh dibaca manual tanpa anotator kedua. Ketiga, hanya *threshold* pada *softmax* maksimum yang diuji. Keempat, *checkpoint* dilatih lebih dulu oleh pengembangnya dengan pembagian data yang tidak terdokumentasi. Pemeriksaan empiris tidak menemukan tanda kebocoran, tetapi jika sebagian data validasi atau uji ikut terlatih, akurasi dan *confidence* pada IndoNLI dapat terlalu tinggi dan kenaikan akurasi karena *abstention* pada IndoNLI dapat terlalu besar; hasil studi kasus tidak terpengaruh karena datanya tidak berasal dari IndoNLI. Penelitian lanjutan perlu menguji kalibrasi ulang, menyetel model pada klaim hoaks dengan anotasi manual termasuk label *neutral*, dan memperbaiki ekstraksi klaim.

5 Kesimpulan

Penelitian ini mengukur efek *abstention* berbasis *threshold* pada model NLI berbahasa Indonesia, dengan *threshold* yang ditetapkan dari *coverage* sasaran pada data validasi untuk tiga *seed*. Pada IndoNLI, sasaran *coverage* 90% menaikkan akurasi pada data yang dijawab sebesar 3,12 ± 0,27 poin pada validasi, 4,32 ± 0,48 poin pada *test_lay*, dan 3,60 ± 0,28 poin pada *test_expert*, dengan *coverage* nyata 84,4% sampai 90,0%. Kenaikan ini konsisten antar *seed* dan membesar ketika *coverage* diturunkan. Pada 758 pasangan klaim (270 hoaks dan 488 valid), akurasi tanpa *abstention* hanya 62,14 ± 9,78%, berkisar 55,41% sampai 73,35% antar *seed*, dan di bawah tebakan kelas mayoritas. Abstention menaikkan akurasi 7,36 ± 1,98 poin pada *coverage* 71,5%, sebanding dengan kenaikan pada *coverage* yang sama di data dalam domain, dan menolak 49% sampai 65% dari hoaks yang diprediksi *entailment*. Namun prediksi *neutral* mencakup 58,5% sampai 89,6% dari kesalahan yang lolos pada hoaks maupun klaim valid, dan 39% sampai 65% dari hoaks yang salah prediksi tetap lolos. Dengan demikian, *abstention* berbasis *confidence* memperbaiki akurasi pada IndoNLI tetapi belum cukup andal untuk dipakai langsung pada verifikasi klaim hoaks nyata, dan pemeriksa manusia tetap diperlukan, apalagi karena variasi antar *seed* yang besar. Pembanding *hypothesis-only* mencapai 59,25 ± 0,31% pada validasi IndoNLI (rata-rata tiga *seed*), dibanding 36,73% pada tebakan mayoritas, sehingga akurasi IndoNLI tidak seluruhnya berasal dari penalaran atas premis. Penelitian lanjutan perlu menguji kalibrasi ulang, menyetel model pada klaim hoaks dengan anotasi manual termasuk label *neutral*, dan memperbaiki ekstraksi klaim.

Referensi

[1] F. Rochman, “Kemkomdigi Identifikasi 1.923 Konten Hoaks Sepanjang 2024,” *ANTARA News Gorontalo*, Jan. 8, 2025. [Online]. Available: https://gorontalo.antaranews.com/berita/292006/kemkomdigi-identifikasi-1923-konten-hoaks sepanjang-2024

[2] D. J. Akhir, “Laporan Mafindo Ungkap Pertamina jadi Sasaran *Hoaks*, ini Daftarnya,” *Okezone*, Oct. 23, 2025. [Online]. Available: https://economy.okezone.com/read/2025/10/23/320/3178861/laporan-mafindo-ungkap-pertamina-jadi-sasaran-hoaks-ini-daftarnya-nbsp

[3] Z. Guo, M. Schlichtkrull, and A. Vlachos, “*A Survey on Automated Fact-Checking*,” *Trans. Assoc. Comput. Linguist.*, Vol. 10, pp. 178–206, 2022, DOI: 10.1162/tacl_a_00454.

[4] T. Schuster, A. Fisch, and R. Barzilay, "*Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence*," in *Proc. 2021 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, 2021, pp. 624–643, DOI: 10.18653/v1/2021.naacl-main.52.

[5] O. Honovich *et al.*, "TRUE: *Re-Evaluating Factual Consistency Evaluation*," in *Proc. 2022 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, WA, USA, 2022, pp. 3905–3920, DOI: 10.18653/v1/2022.naacl-main.287.

[6] R. Mahendra, A. F. Aji, S. Louvan, F. Rahman, and C. Vania, "*IndoNLI: A Natural Language Inference Dataset for Indonesian,"* in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, 2021, pp. 10511–10527, DOI: 10.18653/v1/2021.emnlp-main.821.

[7] J. Fawaid, A. Awalina, R. Y. Krisnabayu, and N. Yudistira, "*Indonesia's Fake News Detection using Transformer Network*," in *Proc. 6th Int. Conf. Sustainable Information Engineering and Technology*, 2021, pp. 247–251, DOI: 10.1145/3479645.3479666.

[8] D. Y. Yefferson, V. Lawijaya, and A. S. Girsang, "*Hybrid model: IndoBERT and Long Short-Term Memory for Detecting Indonesian Hoax News*," *IAES Int. J. Artif. Intell.*, Vol. 13, No. 2, pp. 1913–1924, 2024, DOI: 10.11591/ijai.v13.i2.pp1913-1924.

[9] C. Jocelynne, I. G. N. L. Wijayakusuma, and L. P. I. Harini, "*Detection of Political Hoax News using Fine-Tuning IndoBERT*," *J. Appl. Informatics Comput.*, Vol. 9, No. 2, pp. 354–360, 2025, DOI: 10.30871/jaic.v9i2.8989.

[10] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "*FEVER: A Large-Scale Dataset for Fact Extraction and VERification*," in *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, LA, USA, 2018, pp. 809–819, DOI: 10.18653/v1/N18-1074.

[11] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "*Fact Checking with Insufficient Evidence*," *Trans. Assoc. Comput. Linguist.*, Vol. 10, pp. 746–763, 2022, DOI: 10.1162/tacl_a_00486.

[12] J. Xin, R. Tang, Y. Yu, and J. Lin, "*The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing*," in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and 11th Int. Joint Conf. Natural Language Processing (Volume 1: Long Papers)*, Online, 2021, pp. 1040–1051, DOI: 10.18653/v1/2021.acl-long.84.

[13] B. Wen *et al.*, "*Know Your Limits: A Survey of Abstention in Large Language Models*," *Trans. Assoc. Comput. Linguist.*, Vol. 13, pp. 529–556, 2025, DOI: 10.1162/tacl_a_00754.

[14] D. Hendrycks and K. Gimpel, "*A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks*," in *Proc. Int. Conf. Learning Representations*, 2017.

[15] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "*On Calibration of Modern Neural Networks*," in *Proc. 34th Int. Conf. Machine Learning*, Vol. 70, 2017, pp. 1321–1330.

[16] N. Varshney, S. Mishra, and C. Baral, "*Investigating Selective Prediction Approaches Across Several Tasks in IID, OOD, and Adversarial Settings*," in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, 2022, pp. 1995–2002, DOI: 10.18653/v1/2022.findings-acl.158.

[17] M. Mosbach, M. Andriushchenko, and D. Klakow, "*On the Stability of Fine-Tuning BERT: Misconceptions, Explanations, and Strong Baselines*," in *Proc. Int. Conf. Learning Representations*, 2021.

[18] X. Bouthillier *et al.*, "*Accounting for Variance in Machine Learning Benchmar*ks," in *Proc. Machine Learning and Systems*, Vol. 3, 2021, pp. 747–769.

[19] W. Wongso, "*Indonesian-Roberta-Base-Indonli*," Hugging Face model card. [Online]. Available: https://huggingface.co/w11wo/indonesian-roberta-base-indonli [Accessed: Oct. 7, 2026]

[20] R. Arrahmi, I. Budi, and E. Yulianti, "*Improving Indonesian Claim Verification and Explainability via Evidence-Grounded LLM Reasoning*," *Int. J. Intell. Eng. Syst.*, Vol. 19, No. 6, pp. 649–660, 2026, DOI: 10.22266/ijies2026.0630.34.

[21] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme, "*Hypothesis only Baselines in Natural Language Inference*," in *Proc. 7th Joint Conf. Lexical and Computational Semantics*, New Orleans, LA, USA, 2018, pp. 180–191, DOI: 10.18653/v1/S18-2023.
[22] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith, "*Annotation Artifacts in Natural Language Inference Data*," in *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, LA, USA, 2018, pp. 107–112, DOI: 10.18653/v1/N18-2017.
[23] C. Chow, "*On Optimum Recognition Error and Reject Tradeoff*," *IEEE Trans. Inf. Theory*, Vol. 16, No. 1, pp. 41–46, 1970, DOI: 10.1109/TIT.1970.1054406.
[24] Y. Geifman and R. El-Yaniv, "*Selective Classification for Deep Neural Networks*," in *Advances in Neural Information Processing Systems*, Vol. 30, 2017, pp. 4878–4887.
[25] Z. Jiang, J. Araki, H. Ding, and G. Neubig, "*How can We Know When Language Models Know? On the Calibration of Language Models for Question Answering*," *Trans. Assoc. Comput. Linguist.*, Vol. 9, pp. 962–977, 2021, DOI: 10.1162/tacl_a_00407.
[26] S. Feng, W. Shi, Y. Wang, W. Ding, V. Balachandran, and Y. Tsvetkov, "*Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration*," in *Proc. 62nd Annu. Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, 2024, pp. 14664–14690, DOI: 10.18653/v1/2024.acl-long.786.
[27] Masyarakat Anti Fitnah Indonesia (Mafindo), "*TurnBackHoax*." [Online]. Available: https://turnbackhoax.id [Accessed: Oct. 6, 2026]
[28] ANTARA News, "Berita." [Online]. Available: https://www.antaranews.com [Accessed: Oct. 6, 2026]
[29] Y. Liu *et al.*, "*RoBERTa: A Robustly Optimized BERT Pretraining Approach*," *arXiv preprint arXiv:1907.11692*, 2019.
[30] T. Wolf *et al.*, "*Transformers: State-of-the-Art Natural Language Proce*ssing," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*, Online, 2020, pp. 38–45, DOI: 10.18653/v1/2020.emnlp-demos.6.