# AIは熟練者の判断力を鍛え、若手の思考を代替しうるのか：特許弁護士133人・3ヶ月RCT（NBER論文35720）の徹底解剖

## Gemini 4 Argon

2026年11月7日付のPsyPost記事および翌11月8日付のTechFeed記事（なおTechFeedの該当URLは現在直接アクセス不能ですが、表題および日付から同一研究の解説記事と確認されます）が報じた研究は、マサチューセッツ工科大学（MIT）の労働経済学者デビッド・オーター（David Autor）教授とGoogleの研究チームが2026年9月に全米経済研究所（NBER）から公開（同年10月改訂）したワーキングペーパー第35720号『Does AI Assistance Enhance or Erode Expertise? Evidence from a Three-Month Field Experiment in Patent Drafting（AI支援は専門知を強化するか、それとも侵食するか？ 特許明細書作成における3ヶ月間のフィールド実験からの証拠）』を原典としています[1]。米国の知的財産法律事務所11社に所属する現役の特許弁護士133人を対象に、3ヶ月（90日）間にわたって実施されたこの事前登録済みランダム化比較試験（RCT）は、生成AIの業務導入が抱える「目先の業務パフォーマンス向上」と「長期的な人間のスキル形成」の間の緊張関係（Performance-Learning Tension）を実務環境で切り分けて定量化した実証研究です[2]。

本研究が示した中心的な知見は、「AIを使っている最中に最大の恩恵を受ける層（若手）と、AIを取り上げられた後に自身のスキルとして定着させている層（ベテラン）が逆転する」という非対称性です[2]。特許作成支援AIツール「InFlow」を付与された弁護士群は、AIを使用できる特許起案（ドラフティング）課題において、対照群（AI不使用群）を10日時点で0.34標準偏差（SD、p = 0.03）、90日時点で0.38 SD（p = 0.01）上回り、とりわけ実務経験7年未満の若手弁護士で大幅な品質底上げ（10日時点で+0.58 SD、90日時点で+0.60 SD）と作業時間の短縮（10日時点で約18分短縮）を記録しました[4]。しかし、90日間の実験終了時に「AIの使用を禁じた状態（オフライン）」で既存の特許出願書類の欠陥を見抜き修正する「レッドライニング（査読・赤入れ）課題」を全員に課したところ、**AI**群全体の優位性（**+0.32 SD、p = 0.04**）は実務経験7年以上のベテラン弁護士（+0.45 SD、p = 0.02）に集中しており、若手弁護士の平均的なスキル向上は確認されませんでした（-0.03 SD、回帰p = 0.89、置換検定p = 0.93）[4]。さらに若手のスコア分布は「中下位層の激減」と「最下位層および優良層の同時増加」という二極化（Bifurcation）を起こしており、AIが一部の若手には学習を加速させる「跳躍台（Springboard）」として機能した一方、別の若手には思考と実践を代替して成長を妨げる「クッション（Cushion / Hammock）」として作用した可能性が示唆されています[4]。

ただし、この知見を組織の人材戦略やAI導入計画へ適用する際には、原論文が抱える方法論上の制約を冷静に割り引く必要があります[1]。第一に、事前登録時の目標サンプル数（最大350人）に対して実際のランダム化対象は133人にとどまり、最終90日時点のAIなし課題を完遂したのは91人（うち若手は24人）であるため、サブグループ分析の統計的検出力（Statistical Power）が低く、効果量の過大推計（Type Mエラー）のリスクを伴います[2]。第二に、AI使用禁止課題の提出物91件に対する事後監査で15件に「AI使用の疑い」が検出されており、これらを除外した感度分析では人間の採点による全体効果の統計的有意性が消失し、ベテランの優位性も10%有意水準へと弱まります（一方、LLM採点ではベテランの効果が+0.76 SDへ拡大）[1]。第三に、本研究はGoogleが資金提供し、著者

7名中6名がGoogle社員または契約関係者によって構成された査読前のワーキングペーパーであり、検証されたツールも汎用チャットボットではなく特許実務に特化したGoogle Labs製の専用システム（InFlow）である点に留意しなければなりません[1]。

# 実験デザインの全貌：なぜこの研究は従来のAI生産性研究と一線を画すのか

2023年以降の生成AIに関する初期の実証経済学研究——たとえば文章作成（Noy and Zhang, 2023）、カスタマーサポート（Brynjolfsson et al., 2024 / 2025）、ソフトウェア開発（Peng et al., 2023）、経営コンサルティング（Dell'Acqua et al., 2026）など——は、主に「AIを使用している最中の生産性や成果物の質」を測定し、経験の浅い労働者や低スキル層ほどAIによる底上げ効果が大きいという「平準化効果（Leveling Effect）」を一貫して報告してきました[5]。しかし、それらの研究は「AIという補助輪を外した後に、労働者自身の脳内に専門的な判断力が残っているのか」という人的資本形成の問いには直接答えていませんでした[2]。また、AI撤去後の能力を測った先行研究も、高校生や大学生などの初学者（Bastani et al., 2025）や、専門外の新しいプログラミングライブラリを学ぶ単発セッションの実験（Shen and Tamkin, 2026）などが中心でした[5]。

オーター教授らの研究（NBER Working Paper 35720）が注目される理由は、高度な専門職（米国の現役特許弁護士）を対象に、日常業務への**3**ヶ月間の**AI**導入を介入として設定し、「**AI**支援下の作成能力（**Assisted Performance**）」と「**AI**撤去後の独立した専門的判断力（**Unassisted Professional Judgment**）」を分離して測定した点にあります[2]。

## 事前登録とサンプル選抜・ランダム化設計

本研究は2025年5月に米国経済学会（AEA）のRCTレジストリ（登録番号：AEARCTR-0015823、研究題目：*AI & Expertise research with patent lawyers*）に事前登録され、2025年5月から12月にかけて実施されました[2]。対象となったのは、Googleと継続的かつ非排他的な特許業務委託関係を持つ米国の知的財産（IP）専門法律事務所11社（論文内のTable 1では12コホート）です[5]。当初156名の弁護士が募集され、オンボーディングを完了した133名の現役特許弁護士が実験に組み入れられました[1]。参加弁護士には、実験課題に費やした時間に対して通常の時間あたり顧問料（Hourly Billing Rate）が支払われました[1]。

ランダム化は個人単位で行われ、「所属事務所」および「実務経験年数（Seniority）」を層別化変数（Stratification）として、処置群（**AI**利用群）**90**名、対照群（**AI**不使用群）**43**名の「**2**対**1**（約**68%**対**32%**）」の比率で割り付けられました[1]。統計的検出力を最大化する上では1対1の均等割り付けが望ましいものの、参加事務所側が「より多くの所属弁護士に早期からAIツールを使わせたい」と要望したこと、およびGoogle側のプロダクトユーザー獲得目標との兼ね合いから2対1の比率が採用されています[2]。また、研究チームは当初、介入前のベースライン能力テストの実施を希望しましたが、法律事務所側が「参加募集の障壁になる」として難色を示したため、事前テストは省略されました[5]。その代わり、実務経験年数、連絡メールへの返信速度（日数）、課題遂行への自信、研究要件の理解度といった事前属性において両群間に統計的な有意差がないこと（ベースラインの均衡）がTable A1で確認されています[5]。

弁護士のキャリア段階（Seniority）の定義については、主分析において実務経験**7**年未満（**<7 YoE**）を「若手（**Junior**）」、実務経験7年以上（≥7 YoE）を「ベテラン／シニア（Senior / Veteran）」と区分しています[1]。なお、頑健性確認（Robustness Check）として、経験年数の閾値を3年、5年、7年、9年、11年に変更した再推計も行われており、一貫したパターンが確認されています[5]。

## 検証された専用AIツール「InFlow」とトレーニング内容

実験開始時、処置群と対照群の双方に対して、AIの基礎知識に関する非同期のオンライントレーニングモジュール（AIの基本用語、大規模言語モデル［LLM］の概要、プロンプトエンジニアリングの基礎）が提供されました[5]。その上で、処置群の90名にのみ、ツール固有の追加トレーニングと、当時未公開であったGoogle Labs開発の特許明細書作成支援AIツール「InFlow」（のちにGemini Notebookの一部として統合された機能）へのアクセス権が付与されました[2]。対照群の43名は、3ヶ月の実験期間が終了するまでInFlowおよび生成AIライティングツールの使用を控え、従来のリーガルツールのみで業務を行いました[1]。

InFlowは単なる汎用チャット画面ではなく、特許実務のワークフローに合わせて設計された3つの主要機能を備えていました[1]。第一に、発明者のメモや主クレームから従属クレームや実施形態の詳細説明を生成・書き換えする「プロンプトベースのテキスト改変機能」、第二に、審査官や訴訟弁護士など複数の模擬編集者ペルソナ（Editorial Angles / Lenses）からドラフトに対する多角的な批評や改善案を提示する「フィードバック機能」、第三に、明細書内の用語の不統一やクレーム間の論理的破綻を自動検出する「文書内の矛盾・不整合チェック機能」です[1]。

## 3つの評価課題と5次元のブラインド採点ルーブリック

研究チームは、日常業務でのAI利用と並行して、10日目と90日目の2つの時点で計3つの標準化されたベンチマーク課題（Part 1：ラボ課題）を実施しました[2]。また、これとは別に実際のGoogle関連特許業務における成果物をGoogle社内弁護士がブラインド評価するフィールド追跡（Part 2）および業務ログ調査も計画されました[2]。

| 評価フェーズ・課題名 | 実施時期 | AIツールの使用可否 | 課題の具体的内容と制約条件 | 測定目的（評価対象となる能力） | 完遂者数（N） |
|---|---|---|---|---|---|
| 第1回 起案課題（**10-Day Drafting**） | 10日目 | 処置群のみ可（対照群は不可） | 模擬発明資料（発明者メモ、主クレーム、図面、背景ノート）から、5〜7個（事前登録では5〜8個）の従属クレームと主要実施形態の詳細説明を起案。上限2時間・1,000語以内[1]。 | AI導入初期におけるAI支援下の成果物品質と起案スピードの測定[5]。 | 105名（処置群70名［78%］、対照群35名［81%］）[5] |

| | | | | | |
|---|---|---|---|---|---|
| 第**2**回 起案課題（**90-Day Drafting**） | 90日目 | 処置群のみ可（対照群は不可） | 別の模擬発明パッケージを用いた、より難易度の高い特許起案課題。語数上限を2,000語に拡大[1]。 | 3ヶ月間の継続利用後、AI支援下のパフォーマンス格差が拡大するか縮小するかの検証[5]。 | **98**名（処置群67名［74%］、対照群31名［72%］）[5] |
| 査読・修正課題（**90-Day Redlining**） | 90日目 | 全員使用禁止（完全オフライン） | AIが生成した典型的な欠陥（「特許の禁句［Patent Profanity］」による権利範囲の不当な限定、新規性の過剰主張、技術的誤り、文体上の問題）を含む既存の仮想特許出願書類に対し、自力で赤入れ修正（Redline）と専門的コメントを記述[7]。 | AIという補助輪を外した状態で、弁護士自身の頭脳に**内面化された専門的判断力（Internalized Expert Judgment）**の測定[5]。 | **91**名（若手24名、ベテラン67名）[1] |

すべての提出物は、実験の割り付け（処置群か対照群か、誰が書いたか）を知らされていない独立した外部知財法律事務所の熟練特許弁護士6名のパネル（各提出物を2名の評価者が独立採点）によって、1〜5点のリッカート尺度でブラインド採点されました[2]。評価ルーブリックは、①法的執行可能性（**Enforceability & Legal Strength**：特許成立後の権利行使可能性と法的強度）、②技術的正確性（**Technical Accuracy**：発明の技術的構造や動作原理の正確さ）、③戦略的曖昧さ（**Strategic Ambiguity /** 保護範囲のバランス：米国特許商標庁［**USPTO**］の「最も広義の合理的解釈」基準の下で、将来の商業的権利行使範囲を最大化するためにクレーム文言を戦術的に広く柔軟に設計できているか）、④完全性と目的整合性（**Completeness & Alignment**：特許取得目的に対する網羅性）、⑤明瞭さ（Clarity：記述の明確さと論理構成）の5つの専門的次元で構成されています[2]。

特に90日目の「レッドライニング課題」では、特許実務において致命的とされる「パテント・プロファニティ（Patent Profanity：特許の禁句）」――たとえば「本発明において必須である」といった、本来なら広く取れるはずの特許の権利範囲を自ら不必要に狭めてしまう表現――を正確に見抜き、AIに指示するのではなく自分自身の手で適切な条文へと書き直す実行力が試されました[7]。

## 定量結果の深掘り：AI支援下の「底上げ効果」とAI撤去後の「スキル分岐」

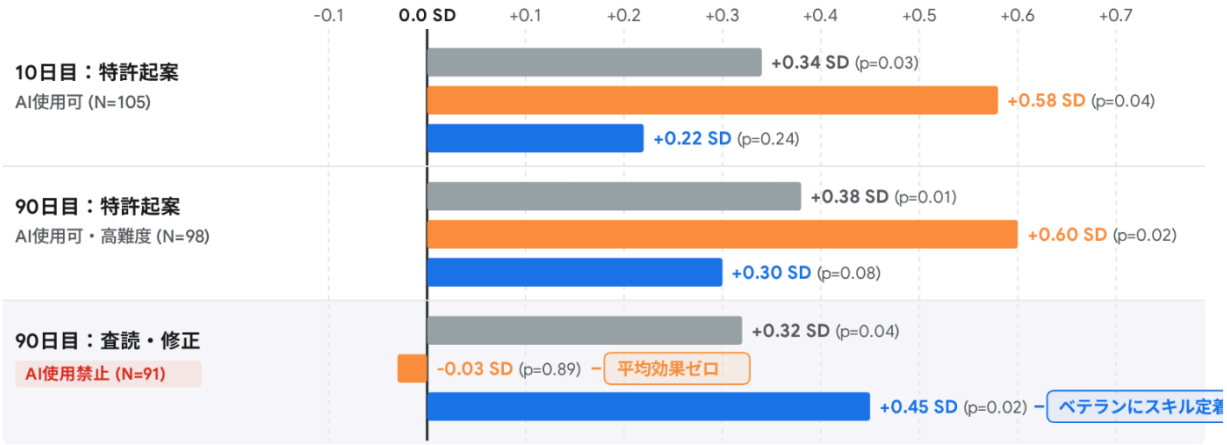
実験から得られた定量データは、AIが「目先の成果物（Output）」に与える影響と、「人間の専門能力（Expertise）」に与える影響がいかに異なるメカニズムで動くかを浮き彫りにしています[4]。

## AI支援下の若手躍進と、ツール撤去後に露呈したベテラン単独のスキル定着と若手の二極化

対象コホート: 若手弁護士（経験7年未満） ベテラン弁護士（経験7年以上） 全体（All Lawyers）

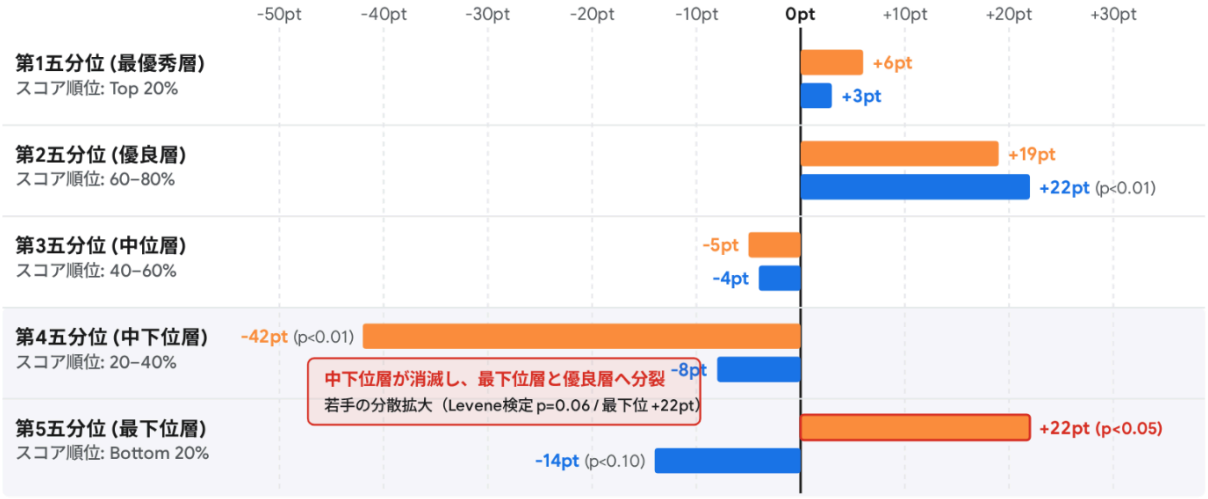
### 1. AI支援下とAI撤去後のパフォーマンス効果量（対照群比 SD）

単位：対照群の標準偏差（SD）｜133名のランダム化比較試験




### 2. AI撤去後の査読課題におけるスコア五分位分布の変化（Table 7：二極化の構造）

単位：各五分位階層に入る確率の変化（%ポイント）｜90日目・AI使用禁止




上段は対照群（AI不使用群）を基準とした処置群（AI利用群）のスコア向上幅（単位：標準偏差 SD）を示す。AIを使用できる10日目・90日目の起案課題では若手（経験7年未満）が**+0.58〜+0.60 SD**と最大の恩恵を受けたが、90日目にAIを禁止して行われた査読・修正（レッドライニング）課題ではベテラン（経験7年以上）のみが**+0.45 SD**の有意な能力向上を示し、若手の平均効果は消失（**-0.03 SD**）した。下段の五分位分布変化（90日目・AIなし査読課題）が示す通り、若手は中下位層（第4五分位）が**-42ポイント**激減する一方で最下位層（第5五分位）が**+22ポイント**急増しており、AIを成長の跳躍台にした層と、思考を代替させて能力を落とした層への深刻な二極化が起きている。

Data sources: NBER Working Paper 35720, MIT Economics (Autor et al.), Google Research Blog, Xiaohu AI Research Summary

## 1. AI支援下の特許起案（10日目・90日目）：「底上げ」による平準化と主観的満足度の高揚

AIツール「InFlow」へのアクセスを与えられた処置群は、10日目の起案課題（完遂105名）で対照群を+0.34 SD（p = 0.03、パーセンタイル順位で10ポイント上昇）上回り、難易度が引き上げられた**90**日目の起案課題（完遂**98**名）でも+0.38 SD（p = 0.01、同11ポイント上昇）と、3ヶ月間にわたって安定した品質向上を達成しました[4]。経験年数別に見ると、若手弁護士（7年未満）の向上幅は10日目で+0.58 SD（p = 0.04）、**90**日目で+0.60 SD（p = 0.02）に達したのに対し、ベテラン弁護士（**7**年以上）の向上幅は**10**日目で+0.22 SD（p = 0.24、非有意）、**90**日目で+0.30 SD（p = 0.08、10%水準で有意）でした[9]。点推定値では若手の改善幅がベテランの約2倍に上りますが、サンプルサイズの制約から両群の改善幅の差（交互作用項）自体は統計的に有意ではありません[1]。

ここで注目すべきは、「AIによる品質向上がスコア分布のどこで起きたか」という点です。論文および**Google Research**の分析によれば、起案品質の向上は「すでに優秀なドラフトをさらに卓越した最高傑作（トップ層）へ引き上げる」ことではなく、「質の低い粗悪なドラフト（下位層）の発生を減らし、中下位・中位層へと底上げする（フロアを引き上げる）」ことによってもたらされました[1]。最上位層（Excellent）の出現頻度はAIを導入しても増加せず、分布の天井（Ceiling）は動かなかったのです[7]。

また、作業時間と主観的評価のデータからは、若手とベテランの間の顕著な行動様式の違いが浮かび上がります[1]。10日目の起案課題において、処置群全体は対照群（平均110分）よりも約10分速く（p = 0.05）課題を完了しました[1]。点推定値の内訳を見ると、若手弁護士は10日目で18分短縮（対照群の平均124分に対して短縮）、90日目でも**22**分短縮したのに対し、ベテラン弁護士の時間短縮は10日目・90日目ともに**5**分にとどまりました（若手とベテランの時間短縮幅の差自体は統計的に非有意）[5]。つまり、若手はAIを使って「時間を節約しつつ一定の質を確保する」行動をとったのに対し、ベテランは浮いた時間を「ドラフトの品質をさらに練り上げる作業」へと充てたことがうかがえます[1]。

さらに事後アンケートにおいて、処置群は強い「スピード向上感」（10日目で+0.72 SD、90日目で+1.14 SD、ともにp < 0.01）を報告しました[5]。とりわけ若手弁護士は10日時点で自らの「成果物の質が上がった」と強く認識し（主観的品質向上 **+0.95 SD**、**p = 0.01**、全体では+0.29 SD、p = 0.23で非有意）、90日時点の特許起案業務に対する職務満足度は+1.48 SD（p < 0.01）という大幅な上昇を示しました（全体では+0.65 SD、p = 0.01、ベテランの満足度変化は+0.28 SD、p = 0.34で非有意であり、若手とベテランの満足度格差は片側p = 0.01で統計的に有意）[5]。若手弁護士は「AIによって自分のスキルが高まり、仕事が速く快適になった」と実感していたものの、その自信はあくまで「AIという外骨格（Exoskeleton）を装着している状態」に依存したものだったことが、続くオフライン課題で明らかになります[1]。

## 2. AI撤去後の査読・修正課題（90日目）：ベテランの躍進と若手の二極化

3ヶ月間の介入終了時、全参加者（完遂91名：若手24名、ベテラン67名）からAIツールを取り上げ、欠陥を含む仮想特許出願書類を手作業で赤入れ修正する「レッドライニング課題」を実施した結果、全体の処置効果は+0.32 SD（p = 0.04、2,000回のモンテカルロ置換検定でも97.1パーセンタイル・p = 0.05、パーセンタイル順位で9ポイント上昇）となりました[5]。しかし、この優位性はベテラン弁護士の+0.45 SD（p = 0.02、置換検定で98.7パーセンタイル・p = 0.02、パーセンタイル順位で13ポイント上昇）によって単独で牽引されており、若手弁護士の平均効果は**-0.03 SD**（回帰**p = 0.89**、置換検定**p = 0.93・46.8**パーセンタイル）と平均的な向上が見られませんでした[5]。

さらに重要なのは、原論文のTable 7で報告された「レッドライニング課題におけるスコア五分位分布

の変化（Distributional Probability Shifts）」です[5]。若手の平均値が「-0.03 SD」であったことは、AIを使った若手が対照群と同じ一様な分布にとどまったことを意味しません[5]。分散の均一性を検定するLevene統計量は、若手群において3.66（p = 0.06）、ベテラン群において7.74（p = 0.01）と有意な分布の変化を示しています[5]。
以下の表は、論文内の主要評価テーブル（Table A：90日目起案課題の専門家人力採点、Table B：同LLM採点、およびTable 7：90日目AIなし査読課題の五分位確率変化）の数値を統合したものです[5]。

| 評価フェーズ・指標・サブスケール | 全体（**All Lawyers**） | 若手弁護士（**<7年**） | ベテラン弁護士（**≥7年**） | 統計的解釈・構造的特徴 |
|---|---|---|---|---|
| **【90日目：AI支援下起案】専門家人力採点（Table A：対照群平均 → 処置群平均）** | *(評価数 N=62 → 137)* | *(評価数 N=18 → 45)* | *(評価数 N=44 → 92)* | ※各評価者レベルの標準正規化平均（括弧内は*SD*）[5] |
| ・総合スコア（Pooled Score） | 0.00 (0.91) → **0.45** (0.71) | -0.32 (0.85) → **0.36** (0.78) | 0.13 (0.92) → **0.50** (0.67) | 若手が対照群のベテラン（0.13）を上回る水準へ上昇[5]。 |
| ・法的執行可能性（Enforceability） | 0.00 (1.00) → **0.35** (0.88) | -0.32 (0.99) → **0.32** (1.01) | 0.13 (0.99) → **0.36** (0.82) | 若手の法的強度スコアが大幅に改善[5]。 |
| ・技術的正確性（Technical Accuracy） | 0.00 (1.00) → **0.46** (0.74) | -0.42 (1.01) → **0.30** (0.77) | 0.17 (0.96) → **0.54** (0.72) | 技術的記述の正確性でも若手の底上げが顕著[5]。 |
| ・戦略的曖昧さ（Strategic Ambiguity） | 0.00 (1.00) → **0.49** (0.85) | -0.26 (0.99) → **0.34** (0.91) | 0.11 (1.00) → **0.56** (0.82) | クレーム範囲の戦術的設定はベテランも大きく伸長[5]。 |
| ・完全性と整合性（Completeness | 0.00 (1.00) → **0.40** (0.78) | -0.19 (0.79) → **0.29** (0.88) | 0.08 (1.07) → **0.46** (0.73) | 抜け漏れ防止において全体およびベテランで |

| | | | | |
|---|---|---|---|---|
| ） | | | | ばらつき（SD）が縮小[5]。 |
| ・明瞭さ（Clarity） | 0.00 (1.00) → **0.56** (0.76) | -0.39 (0.94) → **0.55** (0.76) | 0.16 (0.99) → **0.57** (0.77) | 文章の明瞭さで若手がベテランと同水準に到達[5]。 |
| **【90日目：AI支援下起案】LLM採点（Table B：対照群平均 → 処置群平均）** | *(個体数 N=31 → 67)* | *(個体数 N=9 → 21)* | *(個体数 N=22 → 46)* | ※*LLM*評価者による補完的検証[5] |
| ・総合スコア（Pooled Score） | 0.00 (0.88) → **0.89** (0.92) | 0.41 (0.66) → **0.89** (1.07) | -0.17 (0.91) → **0.89** (0.85) | LLM採点でも処置群全体の改善が確認された[5]。 |
| **【90日目：AI禁止・査読課題】五分位確率変化（Table 7：ITT回帰推計値）** | *(評価数 N=185)* | *(評価数 N=50)* | *(評価数 N=135)* | ※処置群に入ることで各五分位層に属する確率の変化[5] |
| ・Levene分散均一性統計量 | 0.88 (p = 0.35) | **3.66*** (p = 0.06) | **7.74***** (p = 0.01) | 若手・ベテランともに分布形状が統計的に有意に変化[5]。 |
| ・第1五分位（Top / 最優秀層）確率 | +0.05 (SE: 0.06) | +0.06 (SE: 0.04) | +0.03 (SE: 0.07) | 最優秀層の割合は若手・ベテランともにほぼ不変[5]。 |
| ・第2五分位（優良層）確率 | **+0.22***** (SE: 0.06) | +0.19 (SE: 0.15) | **+0.22***** (SE: 0.07) | ベテランは優良層が22ポイント有意に増加（若手も+19pt）[5]。 |
| ・第3五分位（中 | -0.05 (SE: | -0.05 (SE: | -0.04 (SE: | 中位中央層は |

| | | | | |
|---|---|---|---|---|
| 位層）確率 | 0.06) | 0.10) | 0.07) | ほぼ横ばい[5]。 |
| ・第4五分位（中下位・凡庸層）確率 | **-0.17**** (SE: 0.07) | **-0.42***** (SE: 0.16) | -0.08 (SE: 0.08) | 若手の凡庸層が**42ポイント激減（$p < 0.01$）**[5]。 |
| ・第5五分位（Bottom / 最下位層）確率 | -0.05 (SE: 0.06) | **+0.22**** (SE: 0.11) | **-0.14*** (SE: 0.08) | 若手は最下位層が**+22pt急増（$p < 0.05$）**、ベテランは-14pt減少[5]。 |

Table 7の数値は両層の対照的な変化を示しています。ベテラン弁護士の場合、3ヶ月間のAI利用によって第5五分位（最下位層）に落ちる確率が14ポイント減少し（$p < 0.10$）、第4・第3五分位も微減する一方で、第2五分位（優良層）に入る確率が22ポイント増加（$p < 0.01$）するという「分布全体の上方シフト」が起きています[5]。対照的に若手弁護士では、第4五分位（中下位の凡庸なスコア層）が-42ポイント（$p < 0.01$）と大きく空洞化し、その分が第**2**五分位（優良層：**+19**ポイント、標準誤差**0.15**のため非有意）と第5五分位（最下位層：+22ポイント、$p < 0.05$）へと二極化したのです[5]。また、レッドライニング課題に費やした時間は処置群と対照群でほとんど差がなく（処置群の若手は対照群より+4分、ベテランは+1分）、ベテランが平均76分をかけて査読したのに対し、若手は平均69分で終えていました[5]。

# 認知・行動メカニズムの解明：なぜベテランは鍛えられ、若手は二極化したのか

なぜ同じAIツール「InFlow」を3ヶ月間使ったにもかかわらず、ベテランはAIなしの査読能力を高め（+0.45 SD）、若手は優良層へ伸びる者と最下位層へ落ちる者に二極化したのでしょうか。原論文の質的観察記録（Observational note on junior cohort behaviors）およびGoogle Researchの分析は、経験年数によってAIとの関わり方が3つの異なる学習経路へ分岐したことを示しています[5]。

## 1. ベテランが**AI**でさらに鍛えられた理由：基礎的専門知という「フィルター」と高次判断の反復

特許法務やソフトウェア工学などの専門職では、従来「徒弟制度（Apprenticeship）」を通じて暗黙知が継承されてきました[13]。専門家は長年にわたりゼロから初稿を書き、修正し、指導者から赤入れ（フィードバック）を受けるという「認知的摩擦（Friction）」を経験することで、何が優れたクレームであり何が危険な表現（パテント・プロファニティ）であるかという強固なメンタルモデルを構築します[5]。実務経験7年以上のベテラン弁護士は、すでにこの「基礎的専門知（Foundational Expertise）」を確立した状態でInFlowを使い始めました[5]。彼らにとって、AIが生成するドラフトや、InFlowの「複数の編集者ペルソナによる批評」「論理的矛盾チェック」は、そのまま受け入れる完成品ではなく、自らの専門的判断力をぶつけて検証・批評するための対象として機能しました[1]。つまり、ベテランはAIに定型的なテキスト生成を委譲することで認知的余裕を生み出し、3ヶ月間の日常業務を通じて「AIが生成したドラフトの権利範囲の広狭（戦略的曖昧さ）や法的リスクを見極めて修正する」という高次の査

読・判断プロセスを高い頻度で反復していたと解釈できます[5]。これが、10日目・90日目の起案課題でベテランが時間をほとんど節約せず（わずか5分短縮）品質向上に時間を投資した事実とも整合し、AI撤去後に+0.45 SDという判断力向上として現れた要因と考えられます[5]。

## 2. 若手を襲った「硬直的な形式主義（Rigid Formalism）」と認知的オフローディング

一方、基礎的なメンタルモデルがまだ形成途上にある経験7年未満の若手弁護士にとって、AIは「外骨格（Exoskeleton）」として即座に起案スコアを+0.58～+0.60 SD引き上げ、作業時間を18～22分短縮させました[5]。しかし、90日目のオフライン査読課題における若手の実際の編集行動（Redlining Edits）を精査した研究チームは、スキルを伸ばせなかった若手群の提出物に「硬直的な形式主義（Rigid Formalism）」と呼ぶべき特徴的な欠陥パターンを確認しました[7]。

第一の特徴は「上から下への逐次的・近視眼的な作業（Top-to-Bottom Sequential Processing）」です。熟練した特許弁護士が明細書を査読する際は、まず特許の骨格であるクレーム（権利範囲の定義）の構造を把握した上で背景技術や詳細説明を整合させますが、AIに依存していた若手は文書の冒頭から順番に読み進め、重要度の低い導入部のテキストの字句修正（コピーエディティング）に過剰な時間を費やし、核心的な特許クレームの検討に至る前に時間を消耗していました[7]。第二の特徴は「表層的な類語置換への固執（Surface-Level Synonym Swapping）」であり、特許の商業的スコープを広げるための構造的な書き換えを行う代わりに、単なる類義語の置き換えといった表面的な修正に終始していました[7]。第三の特徴は「重大な法的欠陥の放置」であり、権利範囲を不当に狭める「パテント・プロファニティ」や新規性の過剰主張といった深刻な欠陥を見つけた場合でも、それを自力でどう書き直せばよいか分からず、未修正のまま放置するか解決できずに終わっていました[7]。

この現象は、認知科学における「認知的オフローディング（Cognitive Offloading：思考の外部委譲）」の枠組みで説明されます[9]。若手弁護士が初稿を作成する際、本来であれば基礎的な起案と推敲の試行錯誤（Learning-by-Doing）を通じて専門的判断力が養われます[5]。しかし、AIアシスタントが最初から一見もっともらしい初稿（Plausible First Draft）を出力してくれるため、「クッション」経路に陥った若手は、問題を自力で解きほぐすプロセスをAIに委ねてしまいました[5]。その結果、AIがある間は体裁の良い書類を素早く作成できて満足度も上昇（+1.48 SD）したものの、AIを取り上げられ「既存のドラフトに含まれる法的ミスを自力で直せ」と求められた場面で、表面的な字句修正しかできない状態に陥ったと考えられます[5]。

同時に、若手の全員が停滞したわけではなく、第2五分位（優良層）に入る確率が+19ポイント増加した側面（「跳躍台」経路）も報告されています[5]。多忙な法律事務所では、上司が若手のドラフトに対して常に即座のフィードバックを行えるとは限りません[5]。一部の若手弁護士は、InFlowが提供する「複数の模擬編集者レンズによる批評」や「矛盾チェック」、そしてAIが出力する文例を、実務上の「スタイルガイドおよび即時フィードバック装置」として能動的に活用し、判断力の獲得を加速させていました[5]。

## 3. 先行研究との整合性：「補助輪」が学習を阻害する条件

本研究の知見は、2025年から2026年にかけて発表された他の領域のAI学習実験とも整合しています[5]。たとえば、約1,000人の高校生を対象とした数学のフィールド実験（Bastani et al., 2025, PNAS）では、ガードレールのないGPT-4チューターを使った生徒は練習問題の成績が**48%**向上したものの、その後AIを取り上げて実施した試験では対照群よりも成績が**17%**低下しました[11]。また、52人の

ソフトウェア開発者が新しいプログラミングライブラリを習得するランダム化実験（Shen and Tamkin, 2026 等）では、AI支援を使ったグループの事後スキルテスト平均点は約50%にとどまり、**AI**なしで学んだグループの67%（効果量 d = 0.74）を下回りました[5]。さらに、1,222人を対象とした短時間の算術・読解タスク実験でも、約10分間のAI利用でヒントではなく「答え」を求めた参加者は、AI撤去後により少ない問題しか解けなくなり、途中で諦める頻度が高まりました[12]。
オーター教授らの特許弁護士研究は、これら学生や単発セッションの実験で指摘されていた「AIによる学習機会の代替リスク」が、専門職が3ヶ月間日常業務でAIを使った場合にも、「経験の浅いジュニア層の二極化（最下位層の増加）」という形で現れうることをフィールド実験で示した点に意義があります[5]。

# 方法論的限界と批判的検証：データが示していること、まだ証明していないこと

メディアの見出し（「AIはベテランを鍛え、若手の思考を奪う」）をそのまま受け取るのではなく、科学的エビデンスとしての強度を正確に評価するためには、原論文の本文および付録、ならびに事前登録文書（AEARCTR-0015823）に記された方法論的限界と脆弱性を検証する必要があります[1]。

## 1. オフライン課題における「AI使用疑い」の監査（91件中15件のフラグ）と感度分析

本研究の重要な設計は「90日目のレッドライニング課題はAI使用禁止（オフライン）で行われた」という点ですが、実験は各法律事務所の弁護士が自らの執務環境で実施したため、物理的にAI利用を完全遮断して監視することはできませんでした[1]。そのため、「処置群（特にベテラン）がルールに反してAIを使い、その結果としてスコアが高くなったのではないか」という懸念（プロトコル不遵守の可能性）が生じます[1]。
研究チームはこの問題に対処するため、LLMを用いたAI生成テキスト評価プロンプトと人間による目視監査を組み合わせ、少しでも疑いがあればフラグを立てる保守的（Over-classification寄り）な基準で全91件の提出物を監査しました[1]。その結果、**91**件中**15**件（約**16.5%**）の提出物に「**AI**を使用した可能性を示す文体的特徴（**Tells**）」が検出されました[1]。
この15件を巡る感度分析（Sensitivity Analysis）の結果には注意が必要です[1]。フラグが立った15件のスコアはフラグなしの提出物よりも高いものの、その差自体は統計的に有意ではなく、また処置群がレッドライニング課題で時間を短縮していない事実（対照群より1〜4分長い）も、広範なAI不正利用がなかったことを支持しています[5]。しかし、これら疑わしい**15**件をサンプルから除外して再推計すると、人間の専門家採点に基づく「全体サンプルでの**AI**群の優位性（**+0.32 SD**）」は一般的な**5%**有意水準を下回り（統計的に非有意となり）、ベテラン弁護士の優位性（**+0.45 SD**）も**5%**水準から**10%**有意水準（**$p < 0.10$**）へと弱まります[1]。一方で、**LLM**を採点者として用いた評価スコア（**LLM Ratings**）において同じ**15**件を除外すると、全体およびベテランの効果量は逆に拡大し、ベテランの優位性は**+0.76 SD**（**$p < 0.01$**）という高い有意性を示します[5]。提出後のテキスト特徴に基づいて事後的にサンプルを除外する操作はランダム化の前提を崩すため感度診断として扱われますが、人間採点における統計的有意性がこれら15件の扱いに影響を受ける点は重要な留保事項です[1]。

## 2. サンプルサイズの不足・途中脱落（Attrition）と検出力の問題

統計学的な検出力（Statistical Power）の観点からも慎重な解釈が求められます[2]。AEAレジストリの事前登録（AEARCTR-0015823）では「最大350名」のサンプルサイズが計画されていましたが、実

際にランダム化されたのはその38%にあたる133名（対照群43名、処置群90名）でした[1]。さらに90日目のレッドライニング課題を完遂したのは**91**名（ランダム化時点から約**32%**が脱落）であり、内訳はベテラン**67**名に対し、若手は**24**名（評価者2名分の延べ観測数でN = 50）にとどまります[1]。
本研究に対する外部の統計的批評（野崎秀吾氏による検出力試算）では、133名（2対1割り付け）のサンプルで0.32～0.38 SDの効果量を有意水準5%で検出する検出力は41～54%程度、さらに経験年数で分割したサブグループでは0.45 SDを検出する検出力が34～47%程度に落ち込むため、低検出力下で統計的に有意となった推定値が真の効果量より過大に算出される「Type Mエラー（Magnitude Inflation：もし真の効果量が0.20～0.30 SDであれば報告値が2.0～3.6倍に膨らむリスク）」への懸念が指摘されています[2]。
同様に、原論文に対する解説記事（Xiaohu AI等）でも強調されている通り、若手24名における「-0.03 SD（p = 0.89 / 0.93）」という結果も、信頼区間がゼロを跨いでいるため、厳密には「今回のサンプルサイズでは若手において平均的な効果があったという証拠は確認されなかった（平均効果なしの可能性を棄却できない）」ことを意味するのであり、「若手の能力低下が統計的に実証された」と解釈することはできません[2]。

## 3. 事前ベースライン測定の欠如と実務フィールド指標（Part 2）の未検出力

前述の通り、参加事務所の意向により実験開始前のベースライン能力テストが実施されなかったため、本研究は個人の事前・事後の変化量を直接比較するのではなく、ランダム化によって事前能力が均衡しているという仮定（Table A1で属性の均衡を確認）に依拠してスキル獲得を推論しています[1]。また、実験タスク（Part 1）とは別に、日常の特許業務における生産性を自己申告の業務ログで追跡したフィールド調査（Part 2）は、募集開始後に導入されたため対象者が68名（ランダム化対象の51%）にとどまりました[2]。この業務ログ分析では、処置群が特許ドラフト1件を完成させるまでの期間が0.76週間短縮（p = 0.35）し、担当特許件数の差が-1.06件（p = 0.18）、業務上の責任感の差が+0.01 SD（p = 0.98）となりましたが、80%検出力での最小検出可能効果（MDE）が「約2.3週間」であったため、日常業務における5つの自己申告指標はいずれも統計的に有意な差を示していません[5]。

## 4. 未解明の異質性、スポンサーの利益相反（COI）、および外的妥当性

さらに原論文の著者自身が限界として明記している通り、本研究の実験デザインは「なぜAIアクセスが一部の若手には学習の跳躍台として機能し、他の若手にはマイナスの結果をもたらしたのか」という個人間の分岐要因（**Unexplained Heterogeneity**）を説明できていません[5]。なお、論文の目次および補足資料（Supplementary Materials Section 1）には処置群参加者への事後定性インタビュー（Qualitative interviews）が含まれていますが、公開された本文抜粋からはその詳細が省略されています[5]。
ガバナンスと外的妥当性の観点からは、3つの背景を押さえておく必要があります。第一に、実験の直接費用はすべてGoogleが負担し、MITの倫理委員会（COUHES）はMITが本研究の実施主体ではないと判断した上で、Googleの調査手続きの下で実験が遂行されました[1]。著者7名のうち5名（Rodchenko, Iscenko, Strand, Pearl, Ferere）は株式報酬を持つGoogleの正社員、1名（Martin）はGoogleとの有償契約者であり、筆頭著者のAutor教授も「Google Technology and Society Visiting Fellow」（およびHewlett財団、NOMIS財団、Schmidt Sciences AI2050、Smith Richardson財団、Stone財団）からの支援を開示しています[2]。第二に、実験で使われたInFlowは特許専用のカスタム環境（現Gemini Notebookの一部）であり、対象者もGoogleと取引のある米国の高度な知財事務所11社に限定されているため、他国の法域や一般企業の法務部門、汎用チャットボットの利用環境へ

そのまま一般化することはできません[1]。第三に、3ヶ月（90日）という介入期間は、特許実務の専門知を築くのに要する数年の歳月に比べれば依然として短い期間です[1]。

# 組織・実務への戦略的インプリケーション：「スキル負債（Skill Debt）」を防ぐ処方箋

方法論的な留保点を踏まえてもなお、NBER論文35720号が浮き彫りにした「AI支援下のアウトプット品質（Assisted Output）と、人間に内面化される専門判断力（Internalized Skill）の乖離」は、法律事務所のみならず、ソフトウェア開発、コンサルティング、医療など知識集約型産業のマネジメントに対して重要な示唆を与えています[7]。

## 1. 短期指標の陰に潜む「スキル負債（Skill Debt）」の可視化

多くの組織は現在、生成AI導入の効果を「AIを使ってどれだけ作業時間が短縮されたか（10〜22分短縮）」「AIを使った成果物の平均品質がどれだけ上がったか（+0.34〜+0.38 SD）」という**AI**支援下の指標だけで評価しがちです[5]。しかし、本研究が示した通り、それらの指標が最も大きく改善し、本人の職務満足度も高まっている若手層（満足度+1.48 SD）の裏側で、「**AI**なしでは重大な欠陥を修正できない最下位層が**+22**ポイント増加する」というスキル形成上の歪みが生じるリスクがあります[5]。本研究を論じた技術系メディア（Phantom Byteの記事）はこれを「貸借対照表に載らないスキル負債（Skill Debt）」と呼び、同記事が引用するソフトウェア開発の事例でも、AI利用によって構文エラーが76%減少・論理バグが60%減少する一方で、特権昇格の脆弱性経路が322%増加・アーキテクチャ設計上の欠陥が153%増加したとして、若手でも見つけられる安価なエラーが減る裏で、実務経験がなければ見抜けない構造的な欠陥が温存される危険性を指摘しています[17]。

## 2. 実務組織が検討すべき3つの実践的アプローチ

「今日の業務を効率化するAIツールが、明日のシニア専門家を育てる育成経路を損なう」という『徒弟制度のボトルネック（Apprenticeship Bottleneck）』を回避するため、本研究の知見からは以下の3つの実務的アプローチが導かれます[7]。

1. 「**AI**あり評価」と「**AI**なし評価（オフライン・レッドライニング）」の併用：社内のAI活用度や人材評価を行う際、AI支援下の納品物だけで若手の能力成長を測るのではなく、本研究の実験構造を参考に「AIの支援を使わずに、既存のドラフトやコードに含まれる実質的な欠陥（パテント・プロファニティや設計上の問題）を自力で赤入れ修正させるオフライン評価」を定期的に組み合わせることが有効です[7]。これにより、各メンバーにおいてAIが「跳躍台」として機能しているか、それとも「クッション」として基礎スキルの定着を妨げているかを早期に把握できます[5]。
2. 「硬直的な形式主義」を防ぐ構造先行型の実務トレーニング：オフライン課題で苦戦した若手が「文書の冒頭から順番に表面的な類語置換を行い、核心的なクレームの欠陥を放置した」という観察結果は、若手指導の要点がどこにあるかを示しています[7]。若手がAIを使う際には、最初から全文の作成と推敲をAI任せにするのではなく、シニアのクレームツリー（権利範囲の骨格）と照らし合わせながら自力でクレームの道筋を再構築させるなど、判断力を鍛えるための意図的な実践機会（認知的摩擦）を業務プロセス内に残す工夫が求められます[7]。
3. **AI**を「初稿の代筆者」から「多角的な批評・フィードバック装置」へ位置づける活用法：同じInFlowを使った若手の中から優良層（第2五分位+19ポイント）へと伸びた層が生まれた背景には、AIが提供する「複数の模擬編集者レンズからのフィードバック」や「不整合チェック」、質の高い文例を学習の補助として活用した可能性が挙げられています[1]。若手のAI活用においては、単に作業を丸投げして思考をオフロードするのではなく、**AI**の批評機能を通じて「なぜそ

の表現が法的に問題なのか」を比較・検証し、最終的な修正は自らの手で実行するという『跳躍台（Springboard）』型の利用習慣を定着させることが、ツールの恩恵を個人の長期的な専門能力へと結びつける鍵となります[5]。

## 引用文献


1. Prior expertise may determine whether AI builds skills or substitutes, https://www.psypost.org/prior-expertise-may-determine-whether-ai-builds-skills-or-substitutes-for-them-field-experiment-suggests/
2. The biggest gainers from AI were the newcomers. When AI was, https://note.com/shugo/n/nd2b60b059e98?hl=en
3. https://techfeed.io/entries/6aefaa6649a7d098b464cd7b
4. Does AI Assistance Enhance or Erode Expertise? Evidence ... - NBER, https://www.nber.org/papers/w35720
5. Does AI Assistance Enhance or Erode Expertise? Evidence from a, https://economics.mit.edu/sites/default/files/2026-10/Does%20AI%20Assistance%20Enhance%20or%20Erode%20Expertise%20Evidence%20from%20a%20Three-Month%20Field%20Experiment%20in%20Patent%20Drafting.pdf
6. https://shapingwork.mit.edu/research/does-ai-assistance-enhance-or-erode-expertise-evidence-from-a-three-month-field-experiment-in-patent-drafting/
7. https://research.google/blog/does-better-work-always-mean-better-workers/
8. AI Training for Employees and the Future of the Middle Tier | AI Review, https://economy.ac/review/2026/10/202610295778
9. 特許弁護士133人の実験：同じAIでベテランは能力向上, https://xiaohu.ai/ja/article/ai-patent-attorneys-expertise/
10. Does AI assistance enhance or erode expertise? Evidence ... - NBER, https://www.nber.org/system/files/working_papers/w35720/w35720.pdf
11. เราจะสร้างคนอย่างไร ในโลกของ AI - Policy Watch, https://policywatch.thaipbs.or.th/article/government-395
12. (PDF) When AI performs, what remains? An integrative review of, https://www.researchgate.net/publication/417181263_When_AI_performs_what_remains_An_integrative_review_of_generative_AI_and_cognitive_capital_in_individuals_and_organizations
13. Google's AI Patent Tool Sharpens Senior Lawyers but Leaves, https://alphasignal.ai/news/google-s-ai-patent-tool-sharpens-senior-lawyers-but-leaves-juniors-behind
14. AIがジュニアコンサルの仕事を代替すると、将来のパートナーはどう, https://axtruct.co.jp/ai-talent/insights/ai-junior-consultant/
15. Does AI Assistance Enhance or Erode Expertise? Evidence from a, https://policycommons.net/artifacts/60839837/does-ai-assistance-enhance-or-erode-expertise-evidence-from-a-three-month-field-experiment-in-patent-drafting/61737810/
16. Prior expertise may determine whether AI builds skills or substitutes, https://www.reddit.com/r/AllAuthorsWelcome/comments/1xq8xnx/prior_expertise_may_determine_whether_ai_builds/

17. AI Skill Debt: Only Seniors Keep the Gain - PhantomByte Articles, https://articles.phantom-byte.com/ai-skill-debt-nber-patent-lawyer-trial-juniors.html
18. CESifo Working Paper No. 13045 - ifo Institut, https://www.ifo.de/sites/default/files/docbase/docs/cesifo1_wp13045.pdf