استكشاف أثر هندسة الأوامر في مهام الاستدلال القانوني*

تأليف: فانغيي يو، لي كوارتي، فرانك شيلدر**
مختبرات طومسون رويترز، كندا والولايات المتحدة
ترجمة: سعيدة كحيل
المجمع الجزائري للغة العربية

ملخص المقال

لقد أفضى استخدام النماذج اللغوية الكبيرة (LLMs) في أساليب الاستدلال الصفري (Zero-shot) أو بتدريب محدود (Few-shot) في ميدان معالجة اللغة الطبيعية، إلى ظهور حقل علمي جديد يعرف بهندسة الأوامر (Prompt Engineering). وقد بينت الدراسات الحديثة أن أوامر سلسلة التفكير (Chain-of-Thought, CoT) تحدث، عند توظيفها، أثرا بينا في مهام عدة كالعمليات الحسابية والاستدلال المبني على الحس المشترك. ويعمد هذا البحث إلى استجلاء أثر تلك الأساليب في الاستدلال القانوني، وذلك عبر تجارب أجريناها على مهمة الاستلزام في مسابقة COLIEE، وهي مسابقة استخراج المعلومات القانونية والاستلزام القانوني المبنية على اختبار نقابة المحامين اليابانية. وقد قمنا بتقويم أداء النماذج وفق منهج الاستدلال الصفري أو بتدريب محدود، ووفق الضبط الدقيق (Fine-tuning)، مع

---

* **العنوان الأصلي للمقال:**

Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13582-13596, Toronto, Canada. Association for Computational Linguistics.

** **أنجز هذا العمل خلال فترة تدريب في مختبرات طومسون رويترز Thomson Reuters.**

الشروح أو من دونها، إلى جانب طائفة من طرائق هندسة الأوامر. وتشير نتائجنا إلى أن أوامر سلسلة التفكير، وكذلك الضبط الدقيق المصحوب بعبارات تفسيرية، يحسنان الأداء بدرجات معتبرة. غير أن أفضل النتائج تتحقق عند الاستناد إلى أساليب استدلال قانوني محددة، من بينها منهج IRAC (Issue, Rule, Application, Conclusion)، المبني على عرض المسألة، وبيان القاعدة القانونية، ثم إسقاطها على الواقعة، والانتهاء إلى الحكم. وفضلا على ذلك، اتضح لنا أن التعلم بتدريب محدود، عندما تستمد أمثلته عن طريق عنقدة بيانات التدريب السابقة (Clustering)، يؤدي أداء ثابتا عالي الجودة في مهام الاستلزام ضمن مسابقة COLIEE فخلال سنوات إنجاز البيانات التي خضعت للفحص. وقد أفضت تجاربنا كذلك إلى تحسين أفضل نتائج مسابقة COLIEE لعام 2021 من 0.7037 إلى 0.8025، وتجاوز أحسن الأنظمة المشاركة في عام 2022 بحيث بلغت دقتها 0.789.

الكلمات الدالة: هندسة الأوامر، الاستدلال القانوني، عنقدة البيانات، النماذج اللغوية.

1. مقدمة

أضحت آليات التلقين القائم على التعليل في التعامل مع النماذج اللغوية الكبيرة [i] مجالا يتزايد الاعتناء به في أبحاث معالجة اللغة الطبيعية. فالنماذج من طراز GPT-3 التابعة لشركة OpenAI، تحرز نتائج مرضية في بعض المهام ذات الطابع العام؛ إذ بلغت نسبة دقتها 81% وفق أسلوب الاستدلال الصفري (Zero-shot)، و82.8% وفق الاستدلال بتدريب محدود (Few-shot)، وذلك في مهام الاستدلال المعتمدة على مجموعات من مسائل الفهم العام لنظام سؤال جواب فيزيائي PhysicalQA (Brown et al., 2020). غير أننا نلاحظ أن هذه النماذج تبدي عجزا بينا عند الانتقال إلى المعطيات المتخصصة ذات الطابع المهني أو القانوني. وتجاوزت البحوث اللاحقة حدود هذه التساؤلات الأولية، فأدخلت جملة من الأساليب الموسومة بهندسة الأوامر (Prompt Engineering)[ii]، تتراوح بين مطالبة النموذج بأن "يفكر خطوة بخطوة" بغية استنباط تعليل تراكمي ينهض بجواب معين

(Kojima et al., 2022)، وبين تسخير دورة تكرارية تعتمد على تعليلات ينتجها النموذج ذاته لرفع قدرته على بناء استدلال أكثر إحكاما وتفصيلا (Zelikman et al., 2022). وقد أثبتت هذه الأساليب ما يمكن وصفه بتحسين يسير في قدرة النموذج على تبرير الإجابة الصحيحة قياسا على الاستفسارات الأولية السابقة.

يهدف بحثنا إلى الوقوف على آثار هذه الأساليب عند تطبيقها على بيانات ذات تخصص بالغ الدقة، أعني بها المجال القانوني المستمد من اختبار نقابة المحامين اليابانية.[1] ونؤطر عملنا ضمن إطار المسابقة السنوية لاستخراج المعلومات القانونية والاستلزام (COLIEE) التابعة لجامعة ألبرتا (Rabelo et al.,2022)، وهي مسابقة تشتمل على مهام مخصصة لاختبار قدرة النماذج على معالجة الفرضيات القانونية استنادا إلى مقالات سياقية.

نبدأ أولا باستطلاع أساليب الاستدلال الصفري، وصولا إلى الاستدلال بأمثلة محدودة باستخدام نماذج لغوية كبيرة مسبقا، وباستعمال أوامر مستمدة من دراسات سابقة (مثل: "فلنفكر خطوة بخطوة" (Kojima et al.,2022)، أو صيغ وضعناها نحن (مثل: "حدّد رجاء، ما إذا كانت الفرضية الآتية صحيحة أم خاطئة بناء على النص الآتي").

ثم نقيس أثر الضبط الدقيق (Fine-tuning) لنموذج لغوي كبير، بغرض استنتاج إجابات ثنائية (True/False) سواء مع تقديم التعليل أو من دونه، وسواء أكان هذا التعليل منتجا آليا أو مستخرجا من النص القانوني المساند. غير أن أفضل نتائج مجموعة اختبار COLIEE لعام 2021 ما برحت تتحقق وفق أسلوب يعتمد على أوامر قانونية صفرية (Zero-shot legal-prompt)، متجاوزة أرقى أداء منشور سابق بنسبة 14.04%. وترتكز هذه الأوامر القانونية على طرائق الاستدلال القانوني[iii] التي تدرس في كليات الحقوق، ومنها منهج المسألة، والقاعدة، والتطبيق، والخلاصة (Issue, Rule, Application, Conclusion) (Burton, 2017)، وعلى القياس نفسه، حقق أسلوب الاستدلال بتدريب محدود (8-Shot)، عندما استمدت أمثلته عبر عنقدة بيانات تدريب[iv] COLIEE، أفضل أداء على بيانات COLIEE لعام 2022، مسجلا تحسنا عاما في الدقة بنسبة 9.5%.

تشير تجاربنا إلى أن أسلوبي الاستدلال بتدريب محدود، والضبط الدقيق المصحوب بالتعليل، يقدمان نتائج جيدة وثابتة عبر مجموعتي الاختبار اللتين استخدمناهما من مسابقة COLIEE للتقويم. أما أسلوبا الاستدلال الصفري (Zero-shot) والضبط الدقيق المعتمد على الوسوم، فيظهران تباينا أوضح في الأداء على مدار السنتين. ويبدو أن أسلوب الاستدلال الصفري المعتمد على التعليل القانوني يحقق أفضل النتائج خلال سنة واحدة فقط، مما يدل على احتمال ميله إلى فرط التكيف (Overfitting) تجاه مجموعة اختبار بعينها، وهو ما يستدعي مزيدا من البحث في هذه الإستراتيجيات الخاصة بهندسة الأوامر.

2. مهمة الاستلزام القانوني

تعد مسابقة COLIEE (Rabelot et al.,2022,) جهدا بحثيا سنويا انطلق منذ سنة 2014، موجها نحو تطوير الدراسات المتعلقة باسترجاع المعلومات القانونية والاستلزام (Entailment). وتعتمد اثنتان من مهام المسابقة على بيانات مستمدة من اختبار نقابة المحامين اليابانية. ويلزم هذا الاختبار المترشحين من القانونيين أن يبتوا فيما إذا كان بيان قانوني معين صحيحا أم باطلا. ولا سبيل إلى الجواب عن السؤال إلا بعد تحديد النصوص القانونية (المواد) الأكثر ارتباطا به. وتعرف هذه المهمة في المسابقة باسم المهمة 3 (Task 3)، وهي مهمة استرجاعية. وبعد تحديد النصوص المساعدة، يترتب على الممتحن أن يحكم على الفرضية (Hypothesis): هل هي صحيحة أم خاطئة، استنادا إلى النصوص المختارة باعتبارها مقدمة (Premise). وتجسد هذه المهمة 4 (Task 4) في مسابقة COLIEE، وهي محور عملنا المتعلق بالاستدلال القانوني.

الفرضية (Hypothesis): إذا زالت أسباب بدء المساعدة، جاز لمحكمة الأسرة إلغاء القرار المتعلق ببدء المساعدة دون طلب من أي طرف.

المقدمة (Premise): المادة 18 (1): إذا زالت الأسباب المنصوص عليها في الفقرة الأساسية من المادة 1/15، وجب على محكمة الأسرة إلغاء قرار بدء المساعدة، وذلك بناء على طلب الشخص المعني، أو زوجه، أو أحد أقاربه إلى الدرجة الرابعة، أو ولي

القاصر، أو مشرف ولي القاصر، أو المساعد، أو مشرف المساعد، أو النيابة العامة.

(2) وبناء على طلب أحد الأشخاص الوارد ذكرهم في الفقرة السابقة، يجوز لمحكمة الأسرة إلغاء القرار المنصوص عليه في الفقرة (1) من المادة السابقة، كليا أو جزئيا.

الاستلزام (Entailment): لا

وبصياغة أكثر تحديدا، يعرف المنظمون هذه المهمة بوصفها مهمة استلزام قانوني؛ أي:

إذا كان لدينا سؤال قانوني Q، وبعد استرجاع المواد القانونية الملائمة S_1، S_2، ...، S_n، يتعين تحديد ما إذا كانت هذه المواد تستلزم صحة السؤال Q أو تنفيه.

$Entails(S_1, S_2, ..., S_n, Q)$

أو $Entails(S_1, S_2, ..., S_n, \neg Q)$

ويكون جواب هذه المهمة ثنائيا: إما "نعم" (Q) أو "لا" (¬Q). ويُعتمد مقياس الدقة (accuracy) لتقييم الأداء، كما أن خطَّ الأساس العشوائي (أو الاكتفاء دائمًا بالإجابة بـ"نعم" أو دائمًا بـ"لا") يؤدّي إلى مستوى دقة يقارب 0.5، نظرًا إلى أن معظم مجموعات الاختبار تتسم بتوزيع شبه متوازن بين الإجابات الإيجابية والسلبية.

ومن أبرز التحديات التي تواجه هذه المسابقة محدودية حجم بيانات التدريب والاختبار، إذ تتيح الدورات السابقة الأسئلة التي جرى الإجابة عنها في المنافسات المتعاقبة، مما يشكل مجموعا قدره 806 أسئلة. وتضمنت مجموعة اختبار عام 2021 عددا قدره 81 سؤالا، في حين اشتملت مجموعة 2022 على 109 أسئلة. ومن الجدير بالذكر أن البيانات الخاصة بمسابقة COLIEE لا تكون متاحة للباحثين إلا بعد تقديم طلب رسمي إلى الجهة المنظمة للمسابقة.

3. الدراسات السابقة

يمكن الاطلاع على عرض شامل لمهام COLIEE كافة، وما يقابلها من مناهج تقويم ومعالجة، في (Rabelo et al., 2022). وقد اعتمدت أغلب الأنظمة التي تناولت المهمة الرابعة (Task 4) على نماذج مبنية على BERT (Devlin et al., 2019)، رغم محدودية حجم مجموعة التدريب. ومن اللافت للانتباه أن أفضل الأنظمة أداء في عام 2021 استفادت من توسيع حجم بيانات التدريب، كما اعتمد تجميع (Ensemble) نماذج مختلفة مبنية على BERT (Yoshioka et al., 2021). بينما اتجهت أنظمة أخرى إلى استخدام طُرز لغوية متنوعة، منها: DistilRoBERTa (Liu et al., 2019)، وLEGAL-BERT (Chalkidis et al., 2020)، وT5 (Raffel et al., 2020)، وElectra(Clark et al., 2020)، كما في (Schilder et al., 2021).

كما استعانت أنظمة بارزة أخرى في السابق بالنصوص اليابانية الأصلية وطورت نظاما معتمدا على القواعد (Rule-based) لتحديد الاستلزام القانوني (Fujita et al., 2021). في حين استند نظام آخر إلى شبكة عصبية بيانية (Graph Neural Network) اعتمدت في تمثيل العقد على تضمينات DistilRoBERTa وLEGAL-BERT (Wehnert et al., 2021).

ولم يدرس أسلوب التعلم بتدريب محدود (Few-shot) سوى نظام واحد. (Schilder et al., 2021)، حيث استخدم نموذج GPT-3 (Brown et al., 2020). وقد بني ذلك النظام على أسلوب بأمثلة محددة لمعالجة مهمتي 3 و4 معا.[2] وقد أظهرت تجاربهم أن GPT-3 إذا استعمل من غير ضبط دقيق إضافي، ومن دون تقديم المواد القانونية اليابانية ضمن الأوامر، لا يقدم أداء جيدا في هذه المهمة، بل إن نتائجه انخفضت إلى مستوى أدنى من الحد العشوائي. وتشير هذه النتائج إلى أن النماذج اللغوية الكبيرة، رغم سعتها، لا تختزن قدرا كافيا من المعرفة بالقانون الياباني، وأن بلوغ أداء أفضل يستلزم طرائق أكثر تقدما في هندسة الأوامر و/أو الضبط الدقيق للنماذج.

كما نستند إلى دراسات سابقة بحثت في جدوى هندسة الأوامر والتعليل بغية

تمكين النماذج اللغوية الكبيرة من حل مهام أكثر تعقيدا، مثل مسائل الاستدلال المعتمد على الحس المشترك أو المسائل الرياضية اللفظية. وقد استقيناها خصوصا من الدراسات الحديثة التي اقترحت أسلوب أوامر سلسلة التفكير (Chain-of-Thought, CoT)، والتي أظهرت تحسنا واضحا في أداء أسلوبي الاستدلال الصفري (ZS) والاستدلال بتدريب محدود (FS) (Wei et al.; Kojima et al., 2022; Zhang et al., 2022).

وأظهرت المناهج التي دمجت التعليل داخل عملية الاستدلال نتائج أفضل مقارنة بأساليب التلقين المعتادة أو الضبط الدقيق التقليدي. وتعتمد هذه المناهج على قدرة النماذج اللغوية الكبيرة على توليد نصوص تعليلية، ليعاد استخدامها في مرحلة لاحقة لتدريب النموذج وضبطه. وعلى خلاف الأساليب السابقة التي اعتمدت على شروح منشأة يدويا (Rajani et al., 2019)، اقترح زيليكمان وآخرون (Zelikman et al.,2022) منهجا ينتج التعليل تلقائيا عبر نموذج GPT-J في مجالات: المسائل الحسابية اللفظية، وأسئلة الحس المشترك (Commonsense QA)، ومسائل الرياضيات العليا. ويحتفظ نظام STaR بهذه التعليلات إذا أعطى النموذج الجواب الصحيح، لينشئ بذلك مجموعة تدريب جديدة تستخدم لاحقا في تدريب النموذج. وقد أدى هذا الأسلوب إلى تحسن ملحوظ يفوق المناهج التي تضبط على الإجابات فقط، وحقق أداء مقاربا لنموذج أكبر بكثير (GPT-3) في مهمة CommonsenseQA.

4. التجارب والنتائج

يقدم هذا القسم عرضا للمقاربات المتعددة التي قمنا بدراستها باستخدام نموذج لغوي كبير (LLM) مثل GPT-3.5 من أجل معالجة مهمة COLIEE. وقد اشتمل بحثنا على الضبط الدقيق الأساسي، والضبط الدقيق القائم على التعليمات، والاستدلال بعدد معين من الأمثلة (n-shot learning)، والاستدلال بتدريب محدود مع التجميع، والتلقين المبني على أساليب الاستدلال القانوني، إضافة إلى أسلوب سلسلة التفكير (Chain-of-Thought prompting)، ومحاولة حل هذه المهمة الثنائية عبر استخدام تفسيرات يقوم النموذج اللغوي نفسه بتوليدها.

وقد أجريت جميع التجارب باستخدام أحدث نموذج متاح من فئة GPT-3.5 لدى OpenAI، وهو نموذج text-davinci-003، الذي يعد أعلى نماذج GPT-3.5 كفاءة بين المحركات الأربعة المتوفرة وقت إجراء التجارب (davinci، curie، babbage، ada).

استخدمنا في تجاربنا، مجموعات بيانات COLIEE لسنتي 2021 و2022. وقد اخترنا الدقة (Accuracy) معيارا للتقييم، بالنظر إلى التوزيع المتقارب بين الوسوم الإيجابية والسلبية في مجموعات الاختبار الخاصة بمسابقة COLIEE.

1.4. الاستدلال الصفري (Zero-shot, ZS)

قمنا بتصميم مجموعة متنوعة من الأوامر وتضمينها في إعداد الاستدلال الصفري (ZS)، بحيث لا تتطلب أي معرفة متخصصة خاصة بالمجال. وفي القسم 5.4 سنعرض صيغة ZS عند استخدام أوامر مبنية على المجال القانوني، تتضمن أساليب الاستدلال القانوني.

ولضمان أن تكون مخرجات GPT-3.5 حتمية، وأن يعاد تنفيذها بالطريقة نفسها عند إدخال المدخلات ذاتها، قمنا بضبط قيمة الحرارة (temperature) في جميع التجارب عند 0. أما بقية معلمات GPT-3.5 فقد تركناها في قيمها الافتراضية (Top P=1، عقوبة التواتر Frequency penalty=0، عقوبة الحضور Presence penalty=0). كما استخدمنا آلية الترميز الجشع (Greedy decoding) في جميع مراحل هذا العمل من أجل تبسيط الإجراءات.

وبالنسبة لإعداد ZS، فإن المدخل الذي نزود به GPT-3.5 يتكون من الأجزاء الآتية: الأمر الإرشادي (instructive prompt) الذي اخترناه بناء على تجارب سابقة باستخدام النسخة الأقدم من نموذج (text-davinci-002) GPT-3، إضافة إلى زوج المقدمة والفرضية (premise-hypothesis) من بيانات COLIEE، ثم عبارة "صحيح أو خطأ؟ True or False". وقد سبق لنا اختبار ثلاثة أوامر مختلفة باستخدام النموذج الأقدم من GPT-3 (Yu et al., 2022)؛ text-davinci-002 كما هو مبين في الجدول 1. وتشير النتائج إلى أن مجرد إضافة كلمة، التالي: "following" لتحديد

موقع المقدمة يمكن أن يحسن دقة GPT-3 من 0.7160 إلى 0.7407. أما عندما يصبح الأمر أقل دقة في التعليمات، مثل استبدال عبارة، الفرضية المطروحة: " the given premise" بعبارة القانون المدني الياباني: "Japanese civil code statutes"، فإن الدقة تنخفض من 0.7407 إلى 0.7037. ويلاحظ أن عبارة " the given premise" تشير إلى مواد من القانون المدني الياباني الأكثر ارتباطا بالفرضية المعروضة. كما ينبغي الانتباه إلى أن دقة الفائز في مسابقة COLIEE لعام 2021 بلغت 0.7037، وهي الدقة نفسها التي حققها أسوأ أمر في إعداد ZS. وبذلك، فإن مجرد تزويد نموذج GPT-3 بأمر أكثر تحديدا وملاءمة وتعليما يمكن أن يسمح لنا بتجاوز أداء الفائز في COLIEE 2021 بنسبة 3.70%.

اتبعنا في تجاربنا النتائج التي توصلنا إليها باستخدام نموذج text-davinci-002، واخترنا الأمر الثاني: "يرجى تحديد ما إذا كانت الفرضية التالية صحيحة أم خاطئة استنادا إلى المقدمة المعطاة" (prompt2): Please determine if the following hypothesis is True or False based on the given premise). ليكون الأمر الأساسي في جميع المقاربات التي سنناقشها لاحقا باستخدام نموذج text-davinci-003.

وتظهر النتيجة التجريبية أن النموذج الأحدث text-davinci-003 قدم أداء أدنى من text-davinci-002 في إعداد الاستدلال الصفري (ZS)، إذ حقق دقة تبلغ 0.7160.

الجدول 1: أداء نموذج GPT-3 (text-davinci-002) في إعداد الاستدلال الصفري (ZS) باستخدام أوامر مختلفة على مجموعة اختبار COLIEE لعام 2021. وتظهر النتائج أن تغييرات طفيفة في صياغة الأمر يمكن أن تؤثر بدرجة كبيرة في دقة GPT-3.

المدخل	الأمر	الدقة
(prompt +{الأمر) (premise{المقدمة) (+ {الفرضية})	Prompt1: Please determine if the hypothesis is True or False based on the given premise. (الأمر 1: يرجى تحديد ما إذا كانت الفرضية صحيحة أم خاطئة بناء على المقدمة المعطاة.)	0.7160

(hypothesis +) "True or False"?	Prompt2: Please determine if the following hypothesis is True or False based on the given .premise (الأمر 2: يرجى تحديد ما إذا كانت الفرضية الآتية صحيحة أم خاطئة بناء على المقدمة المعطاة.	0.7407
	Prompt3: Please determine if the following hypothesis is True or False based on the Japanese .civil code statutes (الأمر 3: يرجى تحديد ما إذا كانت الفرضية الآتية صحيحة أم خاطئة استنادا إلى مواد القانون المدني الياباني.)	0.7037

2.4. الاستدلال بتدريب محدود (Few-shot, FS)

أثبت براون وآخرون (Brown et al.,2020) أن أداء الاستدلال بتدريب محدود (FS) في النماذج اللغوية الكبيرة يتفوق على أداء الاستدلال الصفري (ZS) بل وعلى بعض تقنيات الضبط الدقيق المتقدمة (SOTA) في مجموعة متنوعة من المهام. وبناء على ذلك، نقوم بتقييم أداء GPT-3.5 في إعداد FS عبر تزويده بأمثلة من أزواج الفرضيات والإجابات، مأخوذة من مصدرين مختلفين: مصادر خارجية ومصادر داخلية.

1.2.4 الموارد الخارجية (External Resource)

استخدمنا موقعا[3] يتضمن أسئلة اختبار نقابة المحامين التي سبق تقييمها، بوصفه موردا خارجيا. وبما أن أزواج الأسئلة والإجابات الأصلية مكتوبة باللغة اليابانية، فقد استخدمنا النسخة الإنجليزية المترجمة بواسطة Google بوصفها بيانات للاستدلال بتدريب محدود. وقد جمع الموقع أسئلة سابقة محلولة في ثمانية موضوعات مختلفة، وقمنا باختيار سؤال واحد مع إجابته من كل موضوع بصورة عشوائية، لتكوين ثمانية أمثلة توضيحية استخدمت بوصفها إعداد "shot-8" في

تجربتنا.

ويعرض مثال إعداد "shot-8" مع زوج محدد من الفرضية والمقدمة من مجموعة اختبار COLIEE لعام 2021 في الشكل 1.

الشكل 1: إعداد shot-8 مع زوج محدد من الفرضية-المقدمة، حيث إن الأمثلة الثمانية مأخوذة من مدونة مخصصة للتحضير لاختبار نقابة المحامين اليابانية، بينما زوج الفرضية-المقدمة مأخوذ من مجموعة اختبار COLIEE لعام 2021.

2.2.4. الموارد الداخلية: التجميع ثم التعلم بتدريب محدود

نظرا لأن مسابقة COLIEE تحظر على المشاركين استخدام موارد خارجية لتدريب نماذجهم، فقد التزمنا بهذا الشرط وامتنعنا عن توظيف أي موارد خارجية لأغراض التدريب. ومع ذلك، أدركنا قيمة الاستفادة من البيانات التي يوفرها منظمو المسابقة، وذلك لإتاحة نهج يقوم على التجميع (Clustering) ثم التعلّم بتدريب محدود (Few-shot Learning). وتقوم هذه الآلية على تحضير أمثلة قليلة بالاعتماد

على البيانات الداخلية فقط، وفق العملية التي سنعرضها فيما يلي:

1. استخدمنا خوارزمية التجميع KMeans من أجل تقسيم مجموعة التدريب إلى عدة عناقيد، مع مراعاة كل من معامل ظل التشابه (silhouette score) والقيود التقنية المفروضة من قبل GPT-3.5. وبما أن هذا النموذج لا يمكنه معالجة أكثر من 4,000 رمز (tokens) كحد أقصى، فإن عدد الأمثلة (shots) المستخدمة يؤثر مباشرة في عدد الرموز ضمن مدخلات GPT-3.5. لذلك، فرضنا قيدا على عدد معين من الأمثلة، بحيث يكون أقل من 15 مثالا.
2. قمنا داخل كل عنقود بتحديد الكلمات الأكثر تمثيلا له، إذ تعد هذه الكلمات مؤشرات أساسية لخصائص كل عنقود، وتمثل عنصرا محوريا في بناء أمثلة الاستدلال بقليل الأمثلة لاحقا.
3. حددنا داخل مجموعة التدريب الفرضيات التي تتضمن الكلمات أكثر تمثيلا لكل عنقود، وتم استخدام هذه الفرضيات لبناء أمثلة الاستدلال، الأمر الذي أتاح تنفيذ إستراتيجية التعلم بتدريب محدود على نحو أكثر فاعلية.

تشير النتائج أنّه في مجموعة اختبار COLIEE لعام 2021، حقّق استخدام البيانات الخارجية والداخلية بوصفها أمثلة استدلال (shots) مستوى الدقة نفسه (0.7654)، متجاوزا أداء الفائز بمسابقة COLIEE 2021 الذي بلغت دقته 0.7037. أما على مجموعة اختبار COLIEE لعام 2022، فإن استخدام البيانات الداخلية، أي نهج التجميع ثم التعلم بتدريب محدود، يحقق دقة قدرها 0.789، متفوقا على استخدام الموارد الخارجية الذي حقق دقة قدرها 0.7615، بينما يتفوق كلا النهجين على أداء الفائز في COLIEE 2022 الذي بلغ 0.6789.

3.4. الاستدلال الصفري مع سلسلة التفكير (Zero-shot-Chain of Thought, ZS-CoT)

أظهر كوجيمة وآخرون (Kojima et al.,2022) أن النماذج اللغوية الكبيرة يمكنها توليد عمليات استدلال على هيئة سلسلة تفكير (Chain-of-Thought)، وأنها تبدي

قدرات استدلالية متفوقة بمجرد إضافة عبارة، لنفكر خطوة بخطوة: " Let's think step by step" قبل كل إجابة في مهام الاستدلال القائمة على الحس المشترك. بناء على ذلك، قمنا بإجراء تجربة باستخدام GPT-3.5 عبر تلقين من مرحلتين (two-stage prompting)، حيث يستفاد من ناتج المرحلة الأولى بوصفه مدخلا للمرحلة الثانية، وذلك بإضافة العبارتين: "لنفكر خطوة بخطوة " « Let's think step by step » و "بناء عليه تكون الفرضية إما صحيحة أو خاطئة" « Therefore, the hypothesis is True or False ».

وبالرغم من أن الاستدلال الصفري-سلسلة التفكيرZS-CoT يبدو نظريا، منهجا مباشرا، إلا أن دقته تكمن في كون التلقين فيه ينفذ مرتين. ففي المرحلة الأولى، وهي مرحلة استخلاص التعليل، نزود GPT-3.5 بالمدخل الآتي: {prompt} + {premise} + {hypothesis} + {CoT}، حيث يمثل الأمر التعليمي (prompt) الصيغة الثانية المذكورة في القسم 1.4، بينما يمثل CoT عبارة « Let's think step by step »، وذلك لأنها أثبتت أداء أفضل في مهام الاستدلال في الحس المشترك (Kojima et al., 2022). ويقوم GPT-3.5 بناء على هذا المدخل بتوليد عملية استدلال، إلا أن النص الناتج قد يتضمن الإجابة النهائية (True) or (false) وقد لا يتضمنها. أما المرحلة الثانية، وهي مرحلة استخلاص الإجابة، فنزود النموذج بكل من مدخلات المرحلة الأولى ومخرجاتها، إضافة إلى أمر ثان بوصفه محفزا للإجابة لذلك، فإن الفرضية (صحيحة أو خاطئة) هي: Therefore, the hypothesis is. ويمرر هذا النص المحفز إلى GPT-3.5 بوصفه مدخلا فينتج الإجابة الثنائية النهائية.

وعند تطبيق أسلوب ZS-CoT باستخدام نموذج text-davinci-003 على مجموعات اختبار COLIEE، ارتفعت الدقة إلى 0.7284 و0.7523 في اختباري 2021 و2022 على التوالي، وهي نتائج أعلى بكثير من تطبيق الأسلوب ذاته باستخدام text-davinci-002 الذي حقق 0.6296 و0.7064 وفقا للترتيب نفسه.

ومن الجدير بالذكر أن GPT-3.5 قادر على توليد تفسيرات ضمن جميع الأساليب السابقة، بما في ذلك ZS وFS وZS-CoT، بالرغم من أن مخرجات ZS و FS لا تتضمن دائما تفسيرات. ويلاحظ على نحو خاص أنه حين تكون الإجابة المتوقعة

هي True، فإن التفسير غالبا لا يتم توليده.

4.4. ضبط النماذج اللغوية (Fine-tuning) مع التعليلات ودونها

يظهر ضبط النماذج اللغوية عادة أداء قويا على عدة مقاييس (Brown et al.,2020) ويقوم هذا الضبط على تحديث أوزان النموذج اللغوي المدرب مسبقا عبر تدريبه على مجموعة بيانات مشروحة مخصصة للمهمة المراد تنفيذها. وقد قمنا بضبط نموذج GPT-3.5 باستخدام مجموعة التدريب الخاصة بمسابقة COLIEE لعام 2021، مع توقع أن يؤدي النموذج المضبوط أداء أفضل مقارنة باستخدام النموذج المدرب مسبقا بصورة مباشرة.

ولتنفيذ عملية الضبط في GPT-3.5، لا بد من توفير مجموعة من عينات التدريب التي تضم المدخلات المتوقعة إلى جانب الناتج[٧] المقابل لها (completion). وقد قمنا بضبط GPT-3.5 باستخدام مجموعة تدريب COLIEE 2021، حيث استخدم زوج المقدمة-الفرضية بوصفه مدخلا، بينما استخدمت الإجابات بوصفها نواتج (completions). كما قمنا بضبط النموذج بطريقتين مختلفتين من المخرجات: مخرجات ثنائية فقط (True أو False) فقط، ومخرجات ثنائية مصحوبة بتعليلات. وتتضمن هذه التعليلات إما تعليلات يولدها GPT-3.5 نفسه، أو تعليلات زائفة (pseudo-explanations) يتم استخلاصها كما سنبينه لاحقا.

ورغم أن وثائق OpenAI تنص على أنه لا حاجة لوجود أمر (prompt) في مدخلات النموذج عند ضبط GPT-3.5، فإننا اعتمدنا مع ذلك على شكلين من المدخلات: أحدهما من دون أمر، والآخر باستخدام الأمر prompt2 الموضح في القسم 1.4، وذلك من أجل دراسة أثر الأوامر في عملية ضبط GPT-3.5. وبعد عملية الضبط، جرى اختبار النموذج المضبوط على مجموعة اختبار COLIEE لعام 2021، كما هو موضح في الجدول 2. وتظهر النتائج أن ضبط GPT-3.5 مع أوامر إرشادية يتفوق على الضبط من دون أوامر بنسبة 23.99%.

وعند ضبط GPT-3.5 باستخدام كل من الإجابات الثنائية والتعليلات، نعتمد نوعين من التعليلات التي تنشأ باستخدام منهجين مختلفين.

1.4.4. التعليل الزائف (Pseudo-explanation)

نقوم باختيار الجملة أكثر ارتباطا بالفرضية من كل مقدمة، وذلك بوصفها تعليلا زائفا (pseudo-explanation). وعلى نحو أكثر تحديدا، فإننا، لكل زوج من الفرضية-المقدمة، نقوم أولا بتجزئة المقدمة إلى جمل منفصلة، ثم نقوم بترميز كل جملة باستخدام المرمز MPNet encoder (Song et al., 2020) المطبق ضمن الحزمة sentence-transformers 2.2.2، ثم نحسب درجة التشابه الكسيني (cosine similarity) بين كل جملة وفرضية، لنختار في نهاية المطاف الجملة التي لها أعلى درجة تشابه لأنها تمثل التفسير المعتمد. وبما أن الضبط باستخدام الإجابات الثنائية وحدها يظهر دقة أعلى حين يتضمن المدخل أوامر إرشادية، فإننا نزود GPT-3.5 في مرحلة الضبط بالصيغة الآتية : "True or False" + {hypothesis} + {premise} + "{prompt}" بينما يكون الناتج (completion) بالصيغة الآتية: + "Because according to" + {label} {pseudo-explanation} وذلك عند تنفيذ الضبط باستخدام كل من الإجابات الثنائية والتعليلات. ومن خلال هذا الأسلوب، نساعد GPT-3.5 على تحديد الجزء الأكثر صلة من المقدمة كي يبني عليه استجابته النهائية. وعلى غرار المخرجات المستخدمة في ضبط GPT-3.5، فإن المخرجات المتولدة في مرحلة الاستدلال (inference) تتضمن أيضا إجابة ثنائية وتعليلا، بحيث يكون هذا التعليل جملة واحدة من المقدمة تم اختيارها بناء على أعلى درجة تشابه مع الفرضية. ويمكن الاطلاع على نتائج الاستدلال في الجدول 2.

الجدول 2: الدقة التي حققها النموذج GPT-3.5 بعد ضبطه على مجموعة اختبار COLIEE لعام 2021. تم ضبط GPT-3.5 باستخدام مدخلات ومخرجات متنوعة. الأمر الموجود في المدخل هو: « Please determine if the following hypothesis is True or False based on the given premise » يرجى تحديد ما إذا كانت الفرضية الآتية صحيحة أم خاطئة بناءً على المقدّمة المعطاة.

المدخل أثناء الضبط	المخرج أثناء الضبط	الدقة
{المتن (premise)} + {الفرضية (hypothesis)}+ True or False?	{الوسم (label)}	0.6173

{الأمر (prompt)} + {المتن} + {الفرضية} +True or False?	{التصنيف}	0.7654
{الأمر} + {المتن} + {الفرضية} +True or False?	{التصنيف} + Because "according to" + {الشرح الزائف (pseudo-explanation)}	0.7160
{الأمر} + {المتن} + {الفرضية} +True or False?	{الشرح المولد بواسطة GPT-3.5}	0.6667

2.4.4. التعليل المولد بالنموذج اللغوي الكبير

يستند هذا الأسلوب إلى منهج "المستدل ذاتي التعليم" "Self-Taught Reasoner" (STaR bootstrapping) (Zelikman et al., 2022)، الذي يقوم بصورة تكرارية على استخدام التعليلات التي يولدها النموذج اللغوي الكبير لزيادة قدرته على تنفيذ استدلال أكثر تعقيدا. ففي دورة واحدة من أسلوب STaR، يقوم النموذج اللغوي بتوليد تعليل للإجابة عن سؤال معين، وإذا كانت الإجابة التي ينتجها خاطئة، فإن النموذج يولد تعليلا بناء على الإجابة الصحيحة، ثم يضبط تدريبيا اعتمادا على التعليلات التي تؤدي إلى إجابات صحيحة. وبما أن إعادة ضبط GPT-3.5 بصورة متكررة تعد عملية مكلفة، فإننا نقدم بديلا لأسلوب STaR bootstrapping. في منهجنا، نقوم بتلقين GPT-3.5 لتوليد تعليلات لكل من: الفرضية-المقدمة-الإجابة، وذلك بإعطائه الأمر التالي من فضلك اشرح لماذا اخترت الفرضية التالية: Please explain why the following hypothesis is + {label} + based on the given premise. + {premise} + {hypothesis} ثم نستخدم رباعيات الفرضية-المقدمة-الإجابة-التعليل لضبط GPT-3.5، بحيث نزوده في مرحلة الضبط بالمدخل : {hypothesis} + {premise} + {prompt} True or False + ويكون الناتج {explanation} :(completion)، وهو التعليل الذي ولده GPT-3.5 في الخطوة السابقة. ومن أجل خفض التكلفة، قمنا بضبط GPT-3.5 مرة واحدة فقط بدلا من التكرار. وفي مرحلة الاستدلال، ينتج النموذج المضبوط إجابات

ثنائية وتعليل يبدو منطقي؛ والدقة الناتجة موضحة في الجدول 2.

وكما يظهر في الجدول 2، فإن ضبط GPT-3.5 باستخدام التعليل الزائف (pseudo-explanation) يتفوق على ضبط GPT-3.5 باستخدام التعليل الذي يولده النموذج نفسه. وترجع أحد الأسباب المحتملة لذلك إلى أن التعليلات التي يولدها GPT-3.5 ليست صحيحة دائما، وأن الضبط باستخدام تعليلات خاطئة يؤدي إلى انخفاض الدقة. وإضافة إلى ذلك، فإن ضبط GPT-3.5 سواء باستخدام التعليل الزائف أو باستخدام التعليلات التي يولدها النموذج نفسه يقدم أداء أدنى مقارنة بضبط GPT-3.5 باستخدام أوامر إرشادية مع الوسوم فقط. ويشير هذا إلى أن تشجيع GPT-3.5 على «التفكير» و«الاستدلال» بشكل مستقل عبر تزويده بأوامر إرشادية مع تدخل أقل قد يؤدي إلى نتائج أفضل.

5.4. الدوافع القانونية في الأوامر والتلقين (Legal Reasoning Prompts, LRPs)

تفترض الأدبيات عادة أن الأوامر تعمل بوصفها تعليمات مهمة ذات دلالة معنوية، وتتطلب من الخبراء تحديد المهمة المعنية بدقة (Mishra et al., 2022) (Lampinen et al., 2022) وقد أدى إعادة صياغة مهام المعالجة اللغوية الطبيعية (NLP) بأوامر مختلفة إلى زيادة ملحوظة في أداء أسلوبي ZS وFS مقارنة بالنماذج المضبوطة تقليديا (wei et al., 2022),(Le Scao and Rush, 2021; Schick and Schütze, 2021; Sanh et al., 2022). وفي ما يتعلق بالمهمة 4 في COLIEE، تظهر التجارب السابقة أن ZS-CoT وFS وضبط GPT-3.5 جميعها تتفوق على أسلوب ZS. وبناء على ذلك، نستخدم أوامر الاستدلال القانوني (LRPs) لتحسين دقة GPT-3.5 في بيئة ZS.

تشكل «الاستدلال القانوني»، و«التفكير الإبداعي»، و«التحليل النقدي» ركائز ما يعرف بالتفكير على طريقة المحامي، وتسهم في بناء الهوية المهنية القانونية (kift et al.,2011) وقد أشار بورتن (Burton،2017) إلى عدد من المختصرات المستخدمة في تدريس الاستدلال القانوني التقليدي، والتي تمثل في ظاهرها مناهج استدلال يعتمد عليها الخبراء والعاملون في المجال القانوني. فمثلا، يمثل الاختصار

IRAC منهجا يقوم على: Issue، Rule، Application، Conclusion. وتشير Issue إلى تحديد المسألة، أو التفكير في الوقائع والظروف التي أدت إلى مثول الأطراف أمام المحكمة. وتمثل Rule القانون الذي يحكم تلك المسألة. أما Application فهي عملية تطبيق القاعدة على وقائع المسألة، وأخيرا الخاتمة Conclusion وهي استنتاج ما إذا كانت القاعدة تنطبق على الوقائع أم لا. وعليه، نطلب من GPT-3.5 "التفكير مثل محام" عبر تلقينه بالسلسلة الآتية: {prompt} + {approach} + {premise} + « True or False » + {hypothesis} ونستخدم عبارة Please analyze if the hypothesis is True or False according to the given legal reasoning approach بوصفها الأمر (prompt). ويظهر الشكل 2 مثالا لمدخلات ومخرجات GPT-3.5 عند تطبيق أمر الاستدلال القانوني في بيئة ZS.

الجدول 3: أداء GPT-3.5 على مجموعتي اختبار COLIEE لعامي 2021 و2022 عند تطبيق أساليب مختلفة للاستدلال القانوني الصفري. وقد تفوق أفضل أسلوب في عام 2021، TREACC، على الفائز في مسابقة COLIEE لعام 2021 بنسبة 14.04%، بينما تجاوز أفضل أسلوب في عام 2022، TRIAccC، الفائز في مسابقة COLIEE لعام 2022 بنسبة 9.5%.

المنهج	التفاصيل	2021	2022
TREACC	موضوع، قاعدة، تفسير، تحليل، حجج مضادة، خلاصة	0.8025	0.7156
TRIAccC	موضوع، قاعدة، قضايا، تحليل (الحالات، الاستنتاج)، خلاصة	0.7778	0.7431
CLEO	دعوى، قانون، تقييم، نتيجة	0.7531	0.7156
ILAC	مسألة، قانون، تطبيق، خلاصة	0.7531	0.7064
IRAACP	مسألة، قاعدة، تطبيق، تطبيق، خلاصة، سياسة	0.7531	0.7156
IRACDD	مسألة، قاعدة، تحليل، خلاصة، دفاع، تعويضات	0.7531	0.6881

IRRAAC	مسألة، قاعدة، استدلال، تطبيق، تحليل بديل، خلاصة	0.7531	0.6789
RAFADC	قاعدة، سلطات/مراجع قانونية، وقائع، قياس وتمييز، خلاصة	0.7531	0.7339
IRAC	مسألة، قاعدة، تطبيق، خلاصة	0.7531	0.7064

تستند هذه المناهج إلى طرائق الاستدلال القانوني التي لخصها بورتن (Burton، 2017)، ويمكن الاطلاع على تفاصيلها في الجدول 3.

وكما يبينه الجدول، يتفوق أسلوب TREACC على الفائز بمسابقة COLIEE 2021 بنسبة 14.04%، كما يتفوق أسلوب TRIAccC على الفائز بمسابقة COLIEE 2022 بنسبة 9.5%. ويلاحظ أيضا أن نموذج GPT-3.5 كان قادرا على تقديم تفسيرات تبدو متوافقة مع منهج الاستدلال القانوني المستخدم. ويمكن الاطلاع على أمثلة لهذه التفسيرات في الملحق.

ويعرض الشكل 3 مقارنة لمستوى الدقة الذي حققه كل أسلوب عند تطبيقه على GPT-3.5. وقد اعتمدنا في الاستدلال الصفري باستخدام محفزات الاستدلال القانوني (ZS-LRPs) على أفضل منهج أداء في كل من عامي 2021 و2022، وهما TREACC وTRIAccC. أما في منهج التدريب المخصص (Fine-tuning) فقد دربنا النموذج ثم اختبرناه باستخدام بيانات التدريب والاختبار الخاصة بالسنة نفسها كما وفرها منظمو COLIEE. وكما يتضح في الشكل، فإن جميع المناهج تفوقت على الفائز بمسابقة COLIEE 2021، لاسيما منهج الاستدلال القانوني (LRP) الذي تفوق بنسبة 14.04%. أما بالنسبة لبيانات اختبار 2022، فقد تفوقت جميع المناهج على نظام الفوز في تلك السنة، باستثناء منهج التدريب دون تعليل. ويلاحظ أن منهج التعلم بتدريب محدود (Few-shot) سواء باستخدام بيانات داخلية أو خارجية، كان الأكثر استقرارا عبر السنتين. كما أجرينا اختبارات كاي-تربيع (Chi-Squared Tests) لقياس ما إذا كانت أفضل الأساليب أداء تتفوق على أنظمة COLIEE الفائزة بدلالة إحصائية. ففي اختبار 2021، تمكن أسلوب ZS-LRPs من التنبؤ الصحيح بـ 65 عينة، بينما حقق نظام الفوز 57 عينة. وبإجراء الاختبار، لم تسجل دلالة إحصائية

بين النتائج (قيمة $p = 0.0518 > 0.05$)، وذلك في الغالب بسبب صغر حجم العينة (81 فقط)[4]. أما بالنسبة لبيانات COLIEE 2022 (عدد العينات $N = 109$)، فقد أثبت الاختبار أن أفضل منهج لدينا (التعلم المحدد باستخدام البيانات الداخلية) يتفوق على النظام الفائز بدلالة إحصائية واضحة (قيمة $p = 0.0138 < 0.05$).

5. الخلاصة والمناقشة

شهدت السنوات الأخيرة تركيزا متزايدا على تسخير أساليب الاستدلال المعتمدة على المحفزات (reasoning-based prompts) من أجل الارتقاء بدقة الإجابة عن الأسئلة في النماذج اللغوية الكبيرة (Zelikman et al.,2022),(et al.,2019 Rajani) ,(Kojima et al., 2022) وإن كانت أساليب الاستدلال الصفري (Zero-Shot) أو تلك التي تقوم على تدريب محدود (Few-Shot) تظهر أداء قويا في المهام الأساسية أو الأسئلة ذات الطابع التتبعي البسيط، فإن معالجة المجالات المعقدة تستلزم إستراتيجيات استدلالية أكثر عمقا، حيث لا تتحقق الإجابة إلا عبر سلاسل منطقية متتابعة داخل سياقات ظاهرة أو ضمنية. أما الممارسون داخل المجال القانوني على وجه الخصوص، فهم يعتمدون عادة على مناهج استدلالية راسخة، من قبيل: تحديد المسألة (Issue)، ثم القاعدة (Rule)، ثم التطبيق (Application)، ثم الخلاصة (Conclusion). وقد قامت فرضيتنا على أن توجيه النماذج اللغوية الكبيرة بمحفزات وفق هذا النمط، إلى جانب الضبط الدقيق القائم على الاستدلال (explanation-based fine-tuning)، قد يسهم في تحسين قدراتها في التعامل مع مهام الاستدلال المعقدة .

مدخل النموذج (Model Input)

يرجى تحليل ما إذا كانت الفرضية صحيحة أم خاطئة وفق نهج الاستدلال القانوني المعطى. المنهج: المسألة، القاعدة، التطبيق، الخلاصة.

المقدمة (Premise): المادة 509: لا يجوز للملتزم بأي من الالتزامات الآتية أن يتمسك بالمقاصد في مواجهة الدائن؛ ومع ذلك، لا يسري هذا الحظر إذا اكتسب الدائن مطالبة تقابل الالتزام محل النزاع من شخص آخر: (1) التزام بالتعويض عن خسارة أو ضرر

ناجم عن فعل غير مشروع ارتكب بسوء نية؛ أو (2) التزام بالتعويض عن خسارة أو ضرر ناتج عن وفاة أو إصابة شخص (باستثناء ما ورد في البند السابق). الفرضية (Hypothesis): إذا كان شخص ما يحمل مطالبة نقدية بالتعويض عن خسارة أو ضرر ألحقه بالمدين، فلا يجوز للمدين أن يتمسك بالمقاصة في مواجهة هذا الشخص باستخدام المطالبة النقدية مقابل التزام التعويض. صحيح أم خطأ؟
مخرج النموذج (Model Output) خاطئ. تفيد المقدمة أن الملتزم بأي من الالتزامات المذكورة لا يجوز له أن يتمسك بالمقاصة ضد الدائن؛ ومع ذلك، لا يسري هذا المنع إذا اكتسب الدائن مطالبة تقابل الالتزام من شخص آخر. وفي الفرضية، يحمل الشخص مطالبة نقدية بالتعويض عن خسارة أو ضرر ضد المدين. وتندرج هذه المطالبة ضمن التعويض عن خسارة أو ضرر ناتج عن وفاة أو إصابة شخص. وعليه، لا يجوز للمدين أن يتمسك بالمقاصة ضد هذا الشخص باستخدام المطالبة النقدية في مواجهة الالتزام الأصلي بالتعويض.

الشكل 2: مدخلات ومخرجات GPT-3.5 عند تطبيق الاستدلال الصفري باستخدام محفز الاستدلال القانوني (ZS with a LR prompt). وهنا استخدم أسلوب IRAC.

لقد أظهرت دراستنا، استنادا إلى بيانات مسابقات COLIEE الحديثة (Rabelo et al.,2022) ، أن أكبر التحسينات في الدقة تحققت من خلال الاستدلال الصفري (Zero-Shot) مع محفزات مبنية على الاستدلال القانوني في GPT-3.5، وبالأخص باستخدام منهج (Burton 2017) Thesis, Rule, Rule, Application, Conclusion، عند اختبار بيانات عام 2021. غير أن هذه المقاربة لم تظهر الأداء نفسه عند اختبار بيانات عام 2022. وفي المقابل، فقد حقق أسلوب التعلم بتدريب محدود (Few-Shot) باستخدام ثمانية أمثلة (Shot-8) أفضل النتائج على بيانات 2022. وبالرغم من هذا التفاوت، فإن كلا الأسلوبين قد تفوقا بوضوح على أحدث الأنظمة المشاركة في COLIEE، حيث جرى رفع أفضل نتيجة في اختبار 2021 من 0.7037 إلى 0.8025، وأفضل نتيجة في اختبار 2022 من 0.6789 إلى 0.789.

علاوة على ذلك، تبين لنا أن مقاربات أخرى أظهرت أداء تنافسيا مقارنة بأنظمة COLIEE الفائزة، مثل أسلوب الاستدلال الصفري مع السلسلة الفكرية (Zero-Shot CoT) (Zelikman et al.,2022)، الذي تفوق على الفائز في COLIEE

بنسبة 3.5% في عام 2021، وبنسبة 10.81% في عام 2022. وكذلك أسلوب الضبط باستخدام تعليلات زائفة (Pseudo-Explanations) من نص المقدمة (Premise)، الذي حقق دقة بلغت 71.6% على بيانات اختبار 2021. ويستفاد من ذلك أن الأساليب التي تعتمد على تدريب محدود (Few-Shot) تعد أكثر فاعية، لما اتسمت به من أداء جيد ومتناسق على مدار العامين، في حين أن أساليب الاستدلال الصفري مع محفزات قانونية تحقق أداء لافتا في سنة معينة، لكنه قد ينخفض في سنة أخرى. وهذا يدل على أن هذه المقاربات ما تزال بحاجة إلى المزيد من البحث والدراسة. وبالرغم من عدم ثبات أداء أسلوب الاستدلال الصفري بالمحفزات القانونية بين عامي 2021 و2022 (فقد حقق TREACC أعلى دقة في 2021 لكنه جاء رابعا في 2022)، فإن جميع الأساليب القانونية التي اختبرناها تجاوزت الأنظمة الفائزة في COLIEE لكلا العامين.

ورغم أن تحليلنا يكشف عن آفاق واعدة في هندسة الأوامر (Prompt Engineering) للمهام الاستدلالية العليا المعتمدة على النماذج اللغوية الكبيرة، فإن مدى قدرة هذه المحفزات على جعل النموذج "يفكر مثل المحامي" ما يزال غير واضح. وعليه، فإن مزيدا من البحث يبدو ضروريا لتبيين فرص نجاح كل من الضبط القائم على الشرح والمحفزات الاستدلالية في المجال القانوني أو في غيره من المجالات المتخصصة.

6. القيود

مع أن النتائج التي حققها هذا البحث تحمل دلالات واسعة، إلا أننا نلفت الانتباه إلى عدد من القيود التي ينبغي مراعاتها:

1. يقتصر نطاق هذه الدراسة على مهام الاستدلال القانوني باستخدام مهمة COLIEE الرابعة (الاستتباع/Entailment)، والمبنية على امتحان المحاماة الياباني. ولذلك، قد لا تعمم هذه النتائج على مهام استدلال قانوني أخرى، خصوصا في الأنظمة القائمة على القانون العرفي الأنجلوسكسوني .(Common Law).

2. تعتمد مهمة COLIEE الرابعة أساسا على مهمة COLIEE الثالثة (الاسترجاع/ Retrieval)، وقد افترضنا في دراستنا أن عملية استرجاع المواد القانونية ذات الصلة - المستخدمة بوصفها مقدمات (Premises) - هي عملية صحيحة بالكامل وخالية من الخطأ.
3. أجريت التجارب باستخدام نسختين فقط من نماذج GPT، ومن غير الواضح كيف يمكن أن تؤدي نماذج لغوية كبيرة أخرى عند تطبيقها على المهمة نفسها.
4. ركزت الدراسة على أساليب الاستدلال الصفري والاستدلال انطلاقا من تدريب محدود، والضبط على النموذج مع الشرح ومن دونه، إلى جانب إستراتيجيات مختلفة في صياغة المحفزات. غير أن الشرح والمحفزات الاستدلالية يصعب التحكم فيهما؛ فبالنسبة للشرح، اعتمدنا على تفسيرات أنشأها النموذج GPT-3.5 دون معرفة مدى موثوقيتها. أما بالنسبة للمحفزات القانونية، فقد بينا أن الإستراتيجيات القانونية تحسن الأداء، لكن يبقى أثر التصريح المباشر بهذه الإستراتيجيات على طريقة معالجة النموذج للمهمة غير واضح.
5. يعد استخدام العنقدة لبيانات التدريب السابقة بوصفها -أمثلة محدودة- أسلوبا جديدا، غير أنه ليس من الواضح مدى فعاليته مع مجموعات بيانات أخرى أو في مجالات مختلفة. ونحن لا نزعم أن هذا الأسلوب سيحقق نتائج أفضل في مهام أخرى.
6. اقتصرت التجارب على أحدث سنتين من بيانات مهمة COLIEE، وقد لا تمتد قابلية تعميم النتائج إلى سنوات أخرى من المهمة. والأهم من ذلك أن حجم بيانات الاختبار صغير نسبيا، وهو ما ينعكس في محدودية الدلالة الإحصائية للنتائج الخاصة بعام 2021.
7. أجريت التجارب على الترجمات الإنجليزية فقط، ولم تنفذ على النصوص اليابانية الأصلية.
8. تظل شركة OpenAI صاحبة السيطرة على نموذج GPT-3 ونماذجها

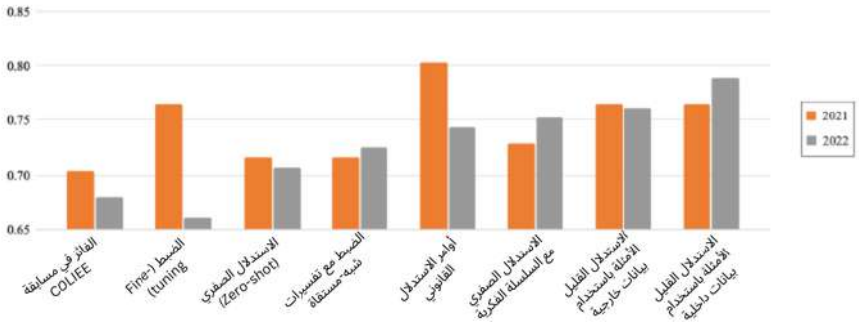
المستقبلية، ولا يمكننا ضمان أن الإصدارات المستخدمة في هذه الدراسة ستكون متاحة للآخرين بالحالة نفسها.

7. بيان أخلاقي

يسعى هذا العمل، في إطار هذه الدراسة، إلى الالتزام بالمعايير الأخلاقية الأساسية المعتمدة في مجال الذكاء الاصطناعي وتعلم الآلة في وضعها الراهن، بما في ذلك (على سبيل المثال لا الحصر) خلو المهمة من أي تحيز معروف تجاه أي فئة أو مجموعة بشرية ضمن نطاق المهمة المدروسة. ومن الطبيعي أن أي اختبار أو تقويم أجري في هذا السياق لا يمكن أن يكون شاملا لجميع الآثار المحتملة، ولا قادرا على رصد كل تأثير ممكن؛ وعليه، ينبغي إجراء تقويم أكثر شمولا في أي أعمال لاحقة مستمدا من هذا البحث.

إن الأساليب والنتائج التي نقترحها، بالرغم من تقديمها بصورة دقيقة وعلى أفضل وجه ممكن، تعتمد على نماذج لغوية مدربة سابقا، وقد ترث بطبيعتها أي تحيزات جوهرية أو أخطاء معرفية كامنة في النموذج الأصلي. وبناء على ذلك، نحن نعرب عن رفضنا الشديد لاستخدام هذه الأساليب في سياقات أو أنظمة قد تفضي إلى آثار مادية ضارة، ومن ذلك على سبيل المثال:

- الإجراءات القانونية أو تقديم الاستشارات، ولاسيما إذا كان ذلك دون مراجعة كافية من قبل محام مرخص مختص في الولاية القضائية المناسبة.
- استخدام النموذج في سياق الامتحانات أو التقييم من غير علم صريح وموافقة مسبقة من جميع الجهات المشرفة على إجراء الامتحان.
- إنتاج أي محتوى موجه للجمهور من دون الإفصاح الصريح بأن هذا المحتوى قد تولد بالكامل بآلة (النموذج اللغوي)، ودون مشاركة بشرية في صياغته.



الشكل 3: مقارنة في مستوى الدقة بين الأنظمة الفائزة في مسابقة COLIEE والنموذج GPT-3.5 عند استخدام أساليب مختلفة.

وعليه، نرى أن الفائدة المثلى من هذا العمل إنما تتحقق عند استخدامه ضمن منهج يقوم على تدخل بشري أثناء التشغيل (Human-in-the-loop)، حيث تتوفر معرفة كافية للتمييز بين المخرجات الصحيحة والمخرجات غير الصالحة التي ينتجها النموذج.

الإحالات

[1] بلغت نسب النجاح في امتحان المحاماة - الذي يُعدّ خطوة حاسمة نحو ممارسة مهنة المحاماة في العديد من البلدان - نحو 80% في الولايات المتحدة سنة 2021 (أنظر : https://tinyurl.com/5yawnh5s)، في حين بلغت 39.2% في اليابان سنة 2020، وهو امتحان يُنظر إليه على نطاق واسع بوصفه أحد أصعب امتحانات المحاماة في العالم (أنظر : https://www.nippon.com/en/ japan-data/h00942/).

[2] قدم المنظمون المهمة 5 التي تتطلب من النظام تنفيذ المهمة 3 (أي: الاسترجاع) والمهمة 4 (أي: الاستلزام) في خطوة واحدة.

[3] https://www.crear-ac.co.jp/shoshi/kakomon/

[4] وقد لاحظنا دلالة إحصائية عند استخدام النسخة القديمة من GPT-3 (text-davinci-002) في منهج الاستدلال الصفري مع محفزات الاستدلال القانوني (ZS with LRPs)، إذ بلغت قيمة $p = 0.0285 < 0.05$.

تعليقات المترجم

[i] النماذج اللغوية الكبيرة (Large Language Models):

تعرف النماذج اللغوية الكبيرة بأنها نماذج حاسوبية متقدمة في مجال معالجة اللغة الطبيعية، تعتمد أساسا على معمارية المحولات (Transformers)، وتدرب على كميات هائلة من البيانات النصية المتنوعة باستخدام تقنيات التعلم العميق، بهدف نمذجة الاحتمالات الإحصائية لتتابعات اللغة الطبيعية. وتتميز هذه النماذج بقدرتها على فهم اللغة البشرية وتوليدها، وأداء مهام لغوية معقدة مثل الترجمة، والتلخيص، والإجابة عن الأسئلة، والتحليل الدلالي، والاستدلال السياقي، وذلك دون تدريب مخصص لكل مهمة (Few-shot Learning،Zero-shot). كما توصف هذه النماذج بـ"الضخمة" نظرا لاحتوائها على مليارات (وأحيانا تريليونات) من المعاملات، وهو ما يسمح لها باكتساب تمثيلات لغوية غنية ومتعددة المستويات، تجمع بين البنية النحوية، والدلالة، والسياق التداولي.

- لتفادي التعدد المصطلحي سنبرر الاختيار الترجمي بمناقشة الفرق بين النماذج كبيرة والنماذج الكبرى وأحيانا الضخمة

(Large Models / Large-Scale Models)

يقصد بـ النماذج الكبيرة النماذج الحسابية، ولا سيما في مجال الذكاء الاصطناعي والتعلم العميق، التي تتميز بـ:

-عدد هائل من المعاملات (Parameters) قد يصل إلى مليارات أو تريليونات،

-تدريبها على كميات ضخمة من البيانات المتنوعة،

-حاجتها إلى بنى حوسبية متقدمة (مثل الحوسبة الموزعة والمعالجات المتخصصة).

يركز هذا المصطلح أساسا على البعد الكمي والحجمي للنموذج، أي على الحجم التقني من حيث البيانات والمعاملات والموارد. مثال: نماذج اللغة واسعة النطاق مثل GPT أو.PaLM

النماذج الكبير(Major Models / Foundation Models)

أما النماذج الكبرى فتحيل، بالإضافة إلى الحجم، إلى المكانة الوظيفية والدور البنوي للنموذج داخل منظومة الذكاء الاصطناعي، إذ تشير إلى نماذج:

(Foundation / Base Models) -مرجعية

-تستخدم كأساس لبناء نماذج أخرى أو لتكييفها مع مهام متعددة،

-تمتلك قدرة تعميم عالية عبر مجالات مختلفة (نصوص، صور، صوت...).

وعليه، فإن وصف "كبرى" لا يقتصر على الحجم، بل يشمل الأهمية، والمركزية، والتأثير البنوي في النظام التقني أو المعرفي.

ينظر في:

Brown, T. B., et al. (2020). Language Models are Few-Shot Learners.Advances in Neural Information Processing Systems (NeurIPS), 33, 1877-190

-الأفضل اعتماد مصطلح النماذج اللغوية الكبيرة لشيوعه في الاستعمال ولتوافقه مع ترجمة اتحاد المجامع العربية.

[ii] هندسة الأوامر (Prompt Engineering) في سياق الذكاء الاصطناعي هي:

عملية تصميم وتشكيل تعليمات (prompts) موجهة إلى نماذج الذكاء الاصطناعي التوليدية، بحيث تعزز جودة ودقة المخرجات وتكيف النموذج لأداء مهام محددة بكفاءة دون تعديل معلمات النموذج نفسه.

بمعنى آخر، هي فن صياغة أوامر واضحة، ودقيقة، وسياقية لتحسين استجابة نماذج اللغة الكبيرة (LLMs) وفق الغرض المطلوب.

مثال عن تقنيات هندسة الأوامر:

الهدف: ضبط مادة من القانون المدني باللغة العربية

-تحديد الدور: خبير قانون مدني.

-توفر السياق القانوني المحدد مثلا: الإحالة على المادة 147 من القانون المدني الجزائري.

-تحديد الغرض المطلوب: ويتم بتوجيه سؤال لبرنامج الترجمة لذكر عناصر الركن المادي والتطبيق القضائي من مرجع موثق باللغة العربية

-- النتائج عادة ما تكون أدق وأكثر فائدة من مجرد "اشرح القانون المدني"، وهذا ما يظهر قيمة هندسة الأوامر في التخصصات الدقيقة مثل القانون.

لمزيد من التوسع في المفهوم التركيب المصطلحي ينظر في:

Comstock, F. (n.d.). Prompt engineering. In Britannica. Retrieved December 26, 2025, from : https://www.britannica.com/technology/prompt-engineering Encyclopedia Britannica

[iii] الاستدلال القانوني:

الاستدلال القانوني عملية عقلية ومنهجية يتم من خلالها الانتقال من القواعد والنصوص القانونية العامة (كالدستور، والقوانين، والاجتهاد القضائي) إلى نتائج خاصة تتعلق بحالة

أو نزاع معين، وذلك اعتمادا على آليات منطقية وتفسيرية مثل القياس، والاستنباط، والاستقراء، وتأويل النصوص، بغرض إصدار حكم قانوني مبرر أو اتخاذ قرار قضائي سليم.

إعادة صياغة من هذا المصدر:

Perelman, Chaïm. (1979)Logique juridique. Nouvelle rhétorique. Paris: Dalloz

[iv] عنقدة بيانات التدريب: تهدف إلى تقسيم بيانات التدريب إلى مجموعات متجانسة تسمى عناقيد(Clusters) ، بحيث تكون البيانات داخل كل عنقود أكثر تشابها فيما بينها من حيث الخصائص أو السمات، وأقل تشابها مع البيانات الموجودة في العناقيد الأخرى. وتستخدم هذه العملية لاكتشاف البنى الكامنة في بيانات التدريب، وتحسين فهم توزيعها، كما تسهم في تحسين أداء نماذج التعلم الآلي من خلال تقليل التعقيد، أو اختيار عينات ممثلة، أو التمهيد لعمليات تصنيف لاحقة.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264-323

ومن حيث صياغة المصطلح فإن الصيغة الصرفية المولدة في الترجمة خاصة في مجال الذكاء الاصطناعي وعلم البيانات تعد أدق من مصطلح التكتل الذي يحيل إلى الدراسات السياسية والاقتصادية رغم القاسم اللغوي المشترك وهو عنقود العنب وكل ما اجتمع على ساق فقد انعقد. وقد استعمل عنقود نجمي في علم الفلك إذا استقامت النجوم.

[v] الإكمال Completion : ما يولده النموذج ردا على الإدخال والمصطلح من مجال الذكاء الاصطناعي ليكون **ناتجا**.

[vi] حق المساطحة :Superficies مصطلح (لاتيني) يستعمل أساسا في القانون الروماني ثم في القوانين المدنية الحديثة، وحق المساطحة ترجمة مستعملة بدقة في الكتابات القانونية العربية الحديثة، خاصة في قوانين الملكية والحقوق العينية. وفي الفقه الإسلامي يستعمل مقابل المصطلح اللاتيني حق القرار. ينظر في هذين المرجعين:

عبد الرزاق السنهوري، الوسيط في شرح القانون المدني، ج 9، دار النهضة العربية، القاهرة.

Gérard Cornu, Vocabulaire juridique, PUF – مادة: Superficie.

قائمة المراجع

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, ClemensWinter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877-1901.

Kelley Burton. 2017. "Think like a lawyer" using a legal reasoning grid and criterion-referenced assessment rubric on irac (issue, rule, application, conclusion). Journal of Learning Design, 10(2):57-68.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2898– 2904, Online. Association for Computational Linguistics.

Kevin Clark, Thang Luong, Quoc V. Le, and Christopher Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In ICLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171-4186. Association for Computational Linguistics.

Masaki Fujita, Naoki Kiyota, and Yoshinobu Kano. 2021. Predicate's argument resolver and entity abstraction for legal question answering: Kis teams at COLIEE 2021 shared task. In Proceedings of the Eighth International

Competition on Legal Information Extraction/ Entailment (COLIEE 2021), pages 15-24.

Sally Kift, Mark Israel, and Rachael Field. 2011. Bachelor of Laws: learning and teaching academic standards statement: December 2010 [Learning and Teaching Academic Standards Project]. Australian Learning and Teaching Council.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. arXiv preprint arXiv:2205.11916.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 537-563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2627–2636, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3470-3487, Dublin, Ireland. Association for Computational Linguistics.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. The Review of Socionetwork Strategies, 16(1):111-133.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1-67.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4932-4942, Florence, Italy. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In International Conference on Learning Representations.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also fewshot learners. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2339-2352, Online. Association for Computational Linguistics.

Frank Schilder, Dhivya Chinnappa, Kanika Madan, Jinane Harmouche, Andrew Vold, Hiroko Bretz, and John Hudzina. 2021. A Pentapus Grapples with Legal Reasoning. In Proceedings of the Eighth International Competition on Legal Information Extraction/ Entailment (COLIEE 2021), pages 60-68.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie- Yan Liu. 2020. MPNet: Masked and permuted pretraining for language understanding. Advances in Neural

Information Processing Systems, 33:16857-16867.

SabineWehnert, Shipra Dureja, Libin Kutty, Viju Sudhi, and Ernesto William De Luca. 2021. Using contextual word embeddings and graph embeddings for legal textual entailment classification. In Proceedings of the Eighth International Competition on Legal Information Extraction/Entailment (COLIEE 2021), pages 69-77.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In International Conference on Learning Representations.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and Maarten Bosma. Chain of thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems.

Masaharu Yoshioka, Yasuhiro Aoki, and Youta Suzuki. 2021. BERT-based ensemble methods with data augmentation for legal textual entailment in COLIEE statute law task. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. ACM.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. arXiv preprint arXiv:2212.01326.

Eric Zelikman, Yuhuai Wu, and Noah D Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. arXiv preprint arXiv:2203.14465.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models.

التعريف بالمؤلفين

- فانغي يو Fangyi Yu

مهندس متخصص في الذكاء الاصطناعي، تعلم آلي في Coursera (يونيو-سبتمبر 2023)، حيث طور نماذج لغوية ضخمة في التنبؤ السلوكي وتحليل البيانات. وهو كذلك لساني تطبيقي ضمن فريق مختبرات طومسون رويترز . مختص في هندسة أوامر النماذج اللغوية الكبيرة، وتقييم الموثوقية والعدالة والتوافق في النماذج الذكية.

مختبرات طومسون رويترز (Thomson Reuters Labs)

19شارع دنكان - تورونتو، أونتاريو M5H 3G6 - كندا

fangyi.yu@tr.com

- لي كوارتي Li Quarti

باحث مشارك في مختبرات طومسون رويترز في مجال الذكاء الاصطناعي وتطبيقاته في القانون.

مختبرات طومسون رويترز (Thomson Reuters Labs)

3 تايمز سكوير - نيويورك، نيويورك 10036 - الولايات المتحدة

lee.quartey@tr.com

- فرانك شيلدر Frank Schilder

مدير بحث في تطوير النماذج اللغوية الذكية وتعلم الآلة، خصوصا في معالجة اللغة الطبيعية (NLP) للحلول القانونية الذكية وتحليل البيانات الكبيرة. يقود فريق من الباحثين والمهندسين في تطوير تقنيات متقدمة في الذكاء الاصطناعي ~~متقدمة~~ تخدم معالجة اللغة القانونية وتطبيقات NLP الذكية.

(Thomson Reuters Labs) مختبرات طومسون رويترز.

طريق أوبرمن - إيغان، مينيسوتا 55123 - الولايات المتحدة 610

frank.schilder@tr.com

الملحق

نورد أدناه نماذج من الشروح التي ولدها نموذج GPT-3.5 (text-davinci-003) عند استعمال أسلوب الاستدلال الصفري (Zero-Shot) وفق أمر TREACC الخاص بالتعليل القانوني، وذلك على مجموعة اختبار COLIEE لعام 2021:

المعرفR02-10-U :

الفرضية:

إن حق المساطحة (Superficies) [vi] هو حق استعمال أرض ما بغية تملك ما، يشيد عليها من أبنية أو ما يغرس فيها من الأشجار أو نبات الخيزران.

النص القانوني:(Premise)

المادة 265: يمنح صاحب حق المساطحة (Superficiary) سلطة استعمال أرض مملوكة للغير، بقصد تملك الأبنية أو الأشجار أو الخيزران المنشأ على تلك الأرض.

التصنيف: الوسم (Label) نعم.

الجواب والتعليل:

صحيح. فحق المساطحة يخول لصاحبه استثمار الأرض من أجل تملك الأبنية أو الأشجار أو الخيزران، وذلك على نحو صريح وارد في المادة 265 A من النظام القانوني. ويعد المستفيد من هذا الحق المسطحي (Superficiary)، وهو المخول بالانتفاع بالأرض، وبإنشاء الأبنية، أو غرس الأشجار والخيزران، وسائر أوجه التصرف التي لا تلحق ضررا بحقوق مالك الأرض. كما يتحمل المسطحي مسؤولية ما قد يترتب من أضرار تلحق بالأرض أو بصاحبها. وبذلك، يتبين أن الفرضية صحيحة.

المعرف R02-13-A:

الفرضية:

هب أن المدين (Obligor A) وضع رهنا على عقاره لصالح الدائن(Obligee B) ، فهل يجوز للدائن (B) أن يحيل ترتيب أولوية الرهن إلى دائن آخر (C) للدائن (A) قبل أن يصبح الأصل المالي للرهن ثابتا ومحددا(Crystallized) ؟

النص القانوني:(Premise)

المادة 398-11 (1): قبل تبلور الأصل المالي للرهن، لا يجوز لصاحب الرهن الدائر (Revolving mortgagee)التصرف فيه وفق أحكام المادة 1/376؛ على أن ذلك لا يمنعه من استعمال الرهن لتأمين مطالبات أخرى.

(2) لا تطبق أحكام المادة 2/377 على المدفوعات الجارية قبل تبلور الأصل المالي للرهن، إذا وقعت ضمن الحالات المستثناة في الفقرة السابقة.

المادة 376 (1): يجوز لصاحب الرهن أن يجعل الرهن ضمانا لمطالبات أخرى، أو أن يحيل أو يتنازل عن الرهن أو عن ترتيب أولويته لمصلحة دائنين آخرين للدائن نفسه.

(2) إذا ترتب عن هذا التصرف استفادة أكثر من شخص، فإن ترتيب حقوقهم ينعقد وفق التسلسل الزمني للتسجيلات الملحقة بسجل الرهن.

التصنيف: الوسم (Label): لا.

الجواب والتعليل: خطأ. لأنه وفقا للمادة 398-11 (1) من النظام القانوني، لا يجوز لصاحب الرهن الدائر أن يتصرف في الرهن، بمقتضى أحكام المادة 1/376، قبل أن يتبلور الأصل المالي للرهن ويتحدد على وجه نهائي. وعليه، لا يسوغ للدائن (B) إحالة ترتيب أولوية الرهن إلى دائن آخر (C) للمدين (A) قبل اكتمال هذا التبلور. ويزداد هذا الحكم وضوحا بما ورد في المادة 2/377 التي تقرر أن المدفوعات الواقعة قبل تبلور الأصل المالي للرهن لا تعتد بها في الحالات المندرجة ضمن الاستثناء الوارد في الفقرة السابقة. ومن ثم، يتبين أن الفرضية المذكورة غير صحيحة.

المعرفR02-36-E :

الفرضية:

إن الدعوى التي يملكها أحد الشركاء على الشيوع (A) ضد شريك آخر (B) فيما يخص المال المشترك، لا يمكن مباشرتها في مواجهة خلفاء (B) الخاصين.

النص القانوني(Premise) :

المادة 254: يجوز مباشرة الدعوى التي يملكها أحد الشركاء على الشيوع ضد الشركاء الآخرين بخصوص المال المشترك في مواجهة خلفائهم الخاصين.

التصنيف: الوسم (Label) لا.

الجواب والتعليل:

خطأ. فالمادة 254 من القانون المدني تجيز صراحة مباشرة الدعوى التي يملكها أحد الشركاء ضد سائر الشركاء فيما يتصل بالمال المشترك، وذلك في مواجهة خلفائهم الخاصين كذلك. وعليه، فإن الدعوى التي يملكها الشريك (A) على الشريك (B) في مال الشيوع يمكن أن تباشر ضد خلفاء (B) الخاصين؛ إذ إن القانون يقر انتقال الحق إلى من يخلف الشريك الذي توجه إليه الدعوى. ومن ثم، تكون الفرضية المذكورة غير صحيحة.

المعرفI-R02-37 :

الفرضية:

إن إظهار الإرادة بالفسخ لا يجوز الرجوع عنه.

النص القانوني:(Premise)

المادة 540:

(1)إذا كان لأحد الطرفين حق الفسخ بمقتضى أحكام العقد أو بموجب القانون، انعقد الفسخ بمجرد إظهار إرادته للطرف الآخر.

(2)لا يجوز الرجوع عن إظهار الإرادة المشار إليه في الفقرة السابقة.

التصنيف(Label) وسم :نعم.

الجواب والتعليل:

صحيح. فالمادة 2/540 تنص بوضوح على أن إظهار الإرادة بالفسخ لا يقبل الرجوع، وذلك بمجرد صدوره على وجه معتبر قانونا. ومتى نصت المادة صراحة بمنع الرجوع، ثبت الحكم ولا يسوغ العدول عنه. وعليه، تكون الفرضية صحيحة.

المعرف R02-27-O :

الفرضية:

يجوز للشريك أن يطلب قسمة مال الشركة قبل تصفيتها إذا وجدت موافقة أغلبية الشركاء.

النص القانوني:(Premise)

المادة 676 (3): لا يجوز للشريك أن يطلب قسمة مال الشركة قبل التصفية.

التصنيف(Label) وسم :لا.

الجواب والتعليل:
خطأ. فالمادة 3/676 تنص بعبارة قاطعة على أنه لا يجوز للشريك طلب قسمة مال الشركة قبل تصفيتها. والحظر هنا مطلق لا يستثنى منه موافقة الأكثرية ولا إجماع الشركاء؛ ذلك أن النص لم يفتح باب الترخيص لأي استثناء من هذا الحكم. وبناء على هذا الإطلاق، يتعذر على أي شريك، مهما بلغ عدد الموافقين، أن يباشر دعوى القسمة قبل تمام التصفية. وعلى هذا، تكون الفرضية المذكورة غير صحيحة.

المعرف R02-1-I :

الفرضية:

إن صدور قرار بالشروع في المساعدة، إذا طلب من غير الشخص المعني بها، يتطلب موافقة هذا الأخير، إلا إذا رأت محكمة الأسرة ملاءمة ذلك.

النص القانوني (Premise):

المادة 15:

(1) يجوز لمحكمة الأسرة أن تقرر بدء المساعدة بحق شخص لا تكفي قدرته على إدراك حالته بسبب اضطراب نفسي، وذلك بطلب من ذلك الشخص، أو من زوجه، أو أحد أقاربه إلى الدرجة الرابعة، أو وليه، أو مشرف الولي، أو القيم، أو مشرف القيم، أو من النيابة العامة؛ على أن لا يعمل بذلك إذا وجدت أسباب كما ورد في المادة 7 أو في الفقرة الأساسية من المادة 11.

(2) إن إصدار قرار بالشروع في المساعدة بناء على طلب صادر من غير الشخص المعني يتطلب موافقته هو.

التصنيف(Label) وسم :لا.

الجواب والتعليل:

خطأ. فالمادة 2/15 تصرح بلا لبس بأن بدء المساعدة بناء على طلب صادر من غير الشخص المعني لا يكون إلا بموافقته، دون تعليق ذلك على رأي المحكمة في الملاءمة أو الصلاحية. وهذا الفهم تؤيده الفقرة 1/15، وإن أجازت تقديم الطلب من غير الشخص المعني، فإنها لا تجيز إصدار القرار إلا إذا وافق هو بنفسه. وعليه يكون ما ورد في الفرضية مخالفا لصريح النص، ومن ثم فهي غير صحيحة.

المعرف R02-12-A :

الفرضية:

أنشأ المدين (A) رهنا عقاريا لصالح الدائن (B) على عقار مملوك له، وقد تم تسجيل هذا الإنشاء. ويجوز للمدين (A) التمسك بالتقادم المسقط للرهن إذا انقضت 20 سنة منذ أصبح الرهن قابلا للتنفيذ، ولو لم ينقض الدين المضمون به.

النص القانوني:(Premise)

المادة 396: لا ينقضي الرهن بالتقادم في مواجهة المدينين والراهنين، إلا إذا انقضى في الوقت نفسه الدين المضمون به.

التصنيف وسم :(Label) لا.

الجواب والتعليل:

خطأ. فالمادة 396 تقطع بأنه لا ينقضي الرهن بالتقادم إلا إذا زال الدين الذي يؤمنه في آن واحد. وعليه، حتى لو مضت عشرون سنة منذ صار الرهن قابلا للتنفيذ، لا يجوز للمدين (A)أن يتمسك بالتقادم المسقط للرهن ما لم ينقض الدين ذاته. ذلك أن الرهن تابع للدين، وجودا وعدما، ولا يتصور انتهاء أحدهما دون الآخر. وقد يثار رأي يدعو إلى إسقاط الرهن بمضي عشرين سنة ولو لم ينقض الدين، إلا أن النص لا يسند هذا القول. ومن ثم، تكون الفرضية غير صحيحة.

المعرف R02-9-O :

الفرضية:

كان (A) مالكا لأرض تقع في منطقة مخصصة للمصائف، فشرع(B) ، الذي يملك أرضا

مجاورة، ببناء سياج على أرض (A) متجاوزا حدودها دون إذنه، ثم اكتمل البناء بعد سنتين. وفي هذه الحالة لا يجوز لـ (A) مطالبة (B) بإزالة السياج عبر دعوى حفظ الحيازة.

النص القانوني:(Premise)

المادة 201 (1): يجب رفع دعوى حفظ الحيازة أثناء وجود التعرض أو خلال سنة من انتهائه؛ على أنه إذا أضر البناء بالشيء محل الحيازة، وكان قد مرت سنة منذ بدء الإنشاء أو اكتمل البناء، فلا تقبل الدعوى.

التصنيف وسم:(Label) . نعم.

الجواب والتعليل:

خطأ. فالمادة 1/201 تجيز رفع دعوى حفظ الحيازة خلال سنة من زوال التعرض، أو عند حصول الضرر بسبب أعمال البناء، سواء مضت سنة منذ البدء أو اكتمل البناء. وفي الحالة المطروحة، اكتمل بناء السياج بالفعل، وهو فعل مؤثر في الحيازة، مما يمكن (A) من إقامة دعوى حفظ الحيازة ضد (B) للمطالبة بإزالة السياج. وعليه، فإن منع الدعوى لا يثبت في هذه الحالة، وتكون الفرضية غير صحيحة.

المعرفU-16-R02 :

الفرضية:

إذا باع (B) المال (X) وسلمه إلى(D) ، مع علم (D) بأن هبة (X) ستلحق ضررا بدائن (A) وهو (C)، جاز لـ (C) أن يطلب من (B) إبطال الهبة التي تمت بين (A) و.*(B)*

النص القانوني:(Premise) حيثيات

المادة 424 (1): يجوز للدائن أن يطلب من المحكمة فسخ تصرف أجراه المدين مع علمه بأنه يلحق الضرر بالدائن؛ شريطة ألا يكون المنتفع من ذلك التصرف (ويشار إليه في هذا القسم بـ"المستفيد") على علم، وقت إبرام التصرف، بأن الدائن سيتضرر منه.

التصنيف: وسم :(Label) نعم.

الجواب والتعليل:

خطأ. فالمادة 1/424 تشترط لجواز فسخ التصرف أن يكون المستفيد عالما وقت إبرام

التصرف بوقوع الضرر بحق الدائن. فإذا لم يعلم (D)، بوصفه المستفيد، بأن الهبة أو البيع يضر بالدائن (C) عند حصول التصرف، لا يجوز لـ (C) طلب فسخه. وبذلك، تكون الفرضية مطروحة على خلاف صريح النص، فهي خاطئة.

A.لكل عملية تقديم:(Submission)

A1.هل قمت بوصف حدود عملك البحثي؟

✔نعم. (القسم 6)

A2.هل ناقشت أي مخاطر محتملة في عملك؟

✔نعم. (الأقسام 5، 6، 7)

A3.هل يقوم المستخلص والمقدمة بتلخيص الادعاءات الرئيسة للورقة؟

✔نعم. (القسم 1)

A4.هل استعنت بمساعدات الذكاء الاصطناعي أثناء إعداد هذه الورقة؟

✔نعم.

تم استخدام ChatGPT في إعداد المستخلص، وإعادة صياغة لغته، وتقديم صياغة أولية لحدود العمل التي جرى تعديلها لاحقا من قبل المؤلفين البشر.

B.هل استخدمت أو أنشأت مواد علمية؟ المستخدمة:

بيانات COLIEE (مهمة 4، حتى عام 2022) في القسمين 2 و3،

ونموذج GPT-3 (text-davinci-002) و GPT-3.5 (text-davinci-003)في القسم 4.

B1.هل قمت بالاستشهاد بمنشئي المواد المستخدمة؟

✔نعم. (راجع المراجع)

B2.هل ناقشت شروط استخدام أو ترخيص تلك المواد؟

✔نعم. (القسمان 2 و4)

B3.هل ناقشت مدى توافق استخدامك لهذه المواد مع الغرض المحدد لها، إن وجد؟

وبالنسبة للمواد التي أنشأتها، هل ذكرت الغرض منها ومدى إمكانية استخدامها خارج سياق البحث؟

✔نعم. (القسمان 2 و4)

لا توجد مواد منشأة جديدة، لكن جرى التحقق من مطابقة الاستخدام للغرض البحثي وعدم خروج البيانات عن سياقها القانوني، كما لا يجوز استخدام البيانات المشتقة خارج إطار البحث العلمي.

B4. هل ناقشت الإجراءات المتخذة للتحقق من خلو البيانات المستخدمة/المجمعة من أي معلومات تعرف بالأشخاص أو تميزهم بشكل مباشر، أو من أي محتوى مسيء، وكذلك الخطوات المتبعة لحمايتها أو إخفاء هويتها؟

✖لم تتم مناقشة ذلك الأمر داخل الورقة،

ولكن جرى فحص البيانات قبل استخدامها، وتبين عدم احتوائها على أسماء أو أي بيانات تعريف شخصية.(PII)

B5. هل قمت بتقديم توثيق للمواد العلمية المستخدمة، مثل: نطاق المجالات، واللغات، والظواهر اللغوية، والفئات والخصائص السكانية الممثلة، إلخ؟

✖غير قابل للتطبيق.

لم يتم إنشاء أي مواد علمية جديدة.

B6. هل أبلغت عن إحصاءات ذات صلة مثل: عدد الأمثلة، وتفاصيل تقسيم بيانات التدريب/الاختبار/التحقق، إلخ، للبيانات المستخدمة/المنشأة؟

✔نعم.

جرى تقديم وصف لتقسيمات البيانات في القسمين 2 و4، حتى بالنسبة لمجموعات البيانات القياسية، لما يوفره ذلك من سياق ضروري لفهم النتائج التجريبية؛ إذ قد تكون الفروق الطفيفة في الدقة ذات أهمية في مجموعات بيانات كبيرة، بينما قد لا تحمل ذات الدلالة في مجموعات صغيرة.

C. هل أجريت تجارب حاسوبية؟

✔ القسم (4)

C1. هل أبلغت عن عدد معاملات النماذج المستخدمة، والميزانية الحاسوبية الإجمالية (مثل ساعات GPU، والبنية التحتية الحاسوبية؟

✔ استخدمنا واجهة برمجة تطبيقات OpenAI وبنى GPT-3.5 ذات الصلة، ضمن ميزانية اشتراك 120 دولارا شهريا.

C2. هل ناقشت إعداد التجارب، بما في ذلك البحث عن القيم المثلى للمعاملات (Hyperparameters) والقيم الأفضل التي تم العثور عليها؟

✔ القسم 4

C3. هل أبلغت عن إحصاءات وصفية للنتائج :مثل حدود الخطأ، أو إحصاءات تلخيصية، وهل كان واضحا ما إذا كانت القيم تمثل أعلى نتيجة، أو المتوسط، أو نتيجة تشغيل واحدة؟

✔ القسم 4

C4. إذا استخدمت حزما جاهزة (للمعالجة أو التقييم)، هل ذكرت التطبيق المستخدم وإعداداته؟ مثل NLTK :، SpaCy، ROUGE، إلخ.

✔ القسم 4.4، ترميز MPNet

D. هل استعنت بمعلقين بشريين (مثل العاملين في المنصات) أو باحثين مشاركين من البشر؟

✘ لم يتم الإجابة.

D1. هل وفرت النص الكامل للتعليمات المعطاة للمشاركين، بما في ذلك لقطات شاشة، أو إخلاء مسؤولية، أو توضيح أي مخاطر؟

✘ لا توجد إجابة.

D2. هل قدمت معلومات عن طريقة استقدام المشاركين (مثل المنصات، أو الطلاب)، وكيف تم دفع أجورهم، وهل كانت الأجور مناسبة؟

✕لا توجد إجابة.

D3.هل ناقشت كيفية الحصول على موافقة المشاركين الذين استخدمت بياناتهم؟

مثلا، هل تضمنت التعليمات توضيح كيفية استخدام البيانات؟

✕لا توجد إجابة.

D4.هل تم اعتماد بروتوكول جمع البيانات أو إعفاؤه من الموافقة الأخلاقية من لجنة أخلاقيات البحث؟

✕لا توجد إجابة.

D5.هل أبلغت عن الخصائص الديموغرافية أو الجغرافية الأساسية للمشاركين الذين شكلوا مصدر البيانات؟

✕ لا توجد إجابة.