המאמר היומי של מייק: 23.04.25  
 The Broader Spectrum of In-Context Learning

למידת In-Context או ל ICL היא יכולת של מודלים לבצע משימות שלא אומנו עליהם במפורש כאשר הם מקבלים כמה דוגמאות לביצוע משימה זו בפרומפט.

המאמר מציע שינוי מהותי מבחינת התבוננות ב- ICL במקום להתייחס אליו כתופעה מצומצמת של למידת few-shot. המחברים מציגים אותו כמנגנון כללי ורחב של הסתגלות הקשרית שנלמד באופן במהלך ה־pretraining על דאטה סדרתית. לשיטתם, כל ירידה עקבית ב־loss שמקורה במידע קודם בסדרה מהווה מקרה של ICL - בין אם מדובר בזיהוי תבנית תחבירית, coreference resolution או topic continuation. מדובר ביכולת הסתגלות שמתפתחת מתוך האימון עצמו, ולא משהו שנדרש ללמד במפורש.

המסגרת המושגית שנבנית כאן נשענת על הבחנה בין שני מעגלי למידה: "outer loop" המתרחש במהלך האימון, ובו המודל לומד על דפוסי דאטה, ו־"inner loop" שבו מתבצעת הסתגלות של המודל בפועל בזמן הריצה, בתוך האקטיבציות של המודל, על סמך ההקשר המקומי בטקסט. הגדרה זו ממקמת את ICL כהתנהגות הסתגלותית emergent, בדומה לגישות של memory-based meta-learning או meta-RL, אך מותאמת למידול שפה בלתי מפוקח.

### סוגים שונים של ICL שלא נחשבים "קלאסיים"

המחברים מציגים קטגוריות שונות של ICL שלא נכנסות למסגרת הרגילה של few-shot learning. כל אחת מראה איך המודל לומד מהקשר בצורה שונה:

Instructional ICL: כאן המודל לומד משימה רק לפי הוראה כתובה ("תרגם מאנגלית לצרפתית"), בלי דוגמאות. הוא מפרש את הפרומפט ומבצע את הפעולה – כלומר, מתנהג כמו מודל שמתאים את עצמו למטרה לפי טקסט בלבד.

Role-based ICL: כשנותנים למודל רמז על מי הוא אמור להיות (למשל "אתה מתרגם מבריק"), הוא משנה את ההתנהגות בהתאם. הרקע שהוא למד עליו כולל הרבה טקסטים עם תפקידים ודמויות, ולכן הוא יודע "לשחק תפקיד" לפי ההקשר.

Explanation-augmented ICL: כשליד כל דוגמה מוסיפים הסבר, המודל עובד יותר טוב. ההסברים עוזרים לו להבין את החוק או התבנית שמאחורי הדוגמאות, לא רק לשנן את התשובות.

Unsupervised ICL: אפילו כשמראים למודל רק שאלות בלי תשובות, הוא מצליח להבין מה המשימה ולפעמים גם לנחש את התשובה. זה קורה כי הוא מזהה מבנים מוכרים שראה באימון, גם בלי שיהיו תשובות זמינות.

Time Series Extrapolation: המודל מצליח לזהות דפוסים ולהמשיך סדרות של מספרים, גם כשיש כמה טרנדים ביחד (למשל עלייה + מחזוריות). הוא עושה את זה בלי אימון נוסף — רק לפי מה שהוא רואה בקונטקסט.

Meta-ICL: כשהמודל רואה כמה משימות ברצף (כל אחת עם דוגמאות), הוא משתפר לאורך הזמן. זה סימן שהוא מצליח לא רק להבין את המשימות, אלא גם להכליל ולזהות מבנים משותפים ביניהן תוך כדי.

### הקשר לשפה: איך ICL צומח ממבני טקסט טבעיים

המחברים מראים שלמודל יש יכולת להסתגל מתוך הקשר בגלל שהוא ראה הרבה דוגמאות של שפה שבה מבנים חוזרים, סדר, תפקידים, והקשרים משתנים לפי מה שנאמר קודם.

Coreference Resolution: המודל יודע לקשר בין ישויות (למשל "היא" מתייחסת ל-"Alice") לפי מה שהיה קודם במשפטים. לפעמים זה פשוט, ולפעמים זה דורש להבין לוגיקה ופרטים מורכבים — כמו בבנמארק Winograd.

Parallel Structure: כשיש כמה משפטים דומים במבנה, המודל לומד את החוק הכללי שמחבר ביניהם. למשל, אם רואים ש-"Alex אוהב חתולים" ו-"Jordan אוהב כלבים", אפשר להסיק את התבנית ולהשלים משפט חדש בהתאם.

Word-Sense Disambiguation: למילים כמו "bank" יש כמה משמעויות. המודל לומד מתוך ההקשר איזו מהן מתאימה – בדיוק כמו שאנחנו עושים בקריאה.

Subject-Verb Agreement: גם מודלים פשוטים מצליחים להבין התאמה בין נושא לפועל. זה סימן שהמודלים הפנימו חוקים תחביריים, ומשתמשים בהם בזמן הריצה.

Topic Modeling: המודל משנה את סגנון הדיבור והמילים שהוא בוחר לפי הנושא של הקטע. גם אם לא מציינים במפורש את הנושא, הוא קולט את זה לפי הקשר ומשנה את ההתפלגות של התחזיות.

### איך לבדוק הכללה ב־ICL

המחברים מציעים שלושה כיוונים עיקריים לבדוק אם מודל באמת יודע להכליל מתוך הקשר:

מה לומדים: האם המודל יכול ללמוד חוק חדש לגמרי מתוך הדוגמאות בקונטקסט, שלא היה באימון? זו ההבחנה בין שינון לבין הבנה אמיתית.

איך לומדים: האם המודל יודע ללמוד את אותה משימה מכמה צורות שונות? למשל, מדוגמאות, מהוראות, מקוד או מטבלה? זו שאלה על גמישות החשיבה של המודל.

איך מיישמים את מה שלומדים: האם אפשר לקחת חוק שהמודל למד ממספרים ולהחיל אותו על מילים? או להסביר אותו? כאן בודקים האם המודל רק "מבצע", או גם מבין לעומק ומסוגל להכליל בין תחומים.

### סיכום

המאמר מציע הסתכלות חדשה על ICL - לא כטכניקה צרה של few-shot prompting, אלא כיכולת הסתגלות כללית שנלמדת תוך כדי אימון על שפה טבעית. לפי הגישה הזו, המודל לומד לזהות מבנים, משימות, תפקידים וחוקים מתוך הקשר, ומשתמש בזה בזמן הריצה, בלי עדכונים. זה כולל גם דפוסים לשוניים פשוטים כמו התאמת פועל, וגם יכולות מורכבות כמו למידת פונקציות או הבנה של הוראות. הגישה של המחברים מחברת בין עולמות של מידול ומטה-למידה של שפה, וייצוג משימות — ומציעה דרכים חדשות למדוד, להבין ולפתח את היכולות של מודלים גדולים.

https://arxiv.org/abs/2412.03782