Doctrine, Data, and the Death of DuPont

Thomas A. Reichert[1]

[1] Assistant Professor of Law, Simmons Law School, Southern Illinois University Carbondale. J.D., M.B.A., M.Eng., Southern Illinois University Carbondale.

ABSTRACT

For fifty years, courts have claimed to apply a comprehensive thirteen-factor test for trademark confusion. They are lying, or at least deeply mistaken. Using AI-powered analysis of 4,000 decisions, this Article proves what practitioners have long suspected: the test has collapsed to just two factors.

Using a large-language-model to extract scored findings for all thirteen factors from approximately 4,000 TTAB inter partes decisions (2000-2025), the study applied statistical models to predict case outcomes. Mark similarity (Factor 1) and goods/services relatedness (Factor 2) alone achieve 99.37% accuracy. Adding the remaining eleven factors increases accuracy to only 99.79%, which is a mere 0.42-point improvement with no practical significance. More striking still, a simple categorical rule predicting confusion if and only if both factors 1 and 2 favor confusion achieves 99.52% accuracy, outperforming the regression models. Further analysis confirms that most secondary factors either repeat information already captured by the core two factors or contribute nothing meaningful to outcomes.

These findings confirm at scale what prior scholarship has suggested: in determining trademark confusions, courts pay lip service to comprehensive multi-factor analysis while actually deciding cases based on just two considerations. The results also reveal concrete harms from this doctrinal gap: parties spend substantial resources litigating factors that do not influence outcomes, case results become harder to predict in advance, and adjudicators exercise broad discretion without meaningful constraints.

The Article explores how these findings might inform doctrinal reform, how reforms would center the two determinative factors and limit secondary considerations to narrow tiebreakers in genuinely ambiguous cases. Finally, it advances a broader "multifactor collapse" hypothesis and outlines a research agenda for testing whether other legal balancing frameworks exhibit similar patterns where doctrinal complexity masks simpler underlying decision-making.

Table of Contents

PART 1:

DOCTRINE

I. The Birth of the Thirteen Factors

A. In re E.I. DuPont DeNemours & Co. , 476 F.2d 1357 (CCPA 1973)

In 1973, the United States Court of Customs and Patent Appeals faced a challenge: how to systematically decide whether two trademarks are too similar?[2] The court's answer in *In re E.I. DuPont DeNemours & Co.* became the foundation of modern American trademark law, shaping how courts and the United States Patent and Trademark Office ("USPTO") analyze confusion to this day.[3] Rather than trying to craft one universal test, the court took a different approach. As they put it, "[t]here is no litmus rule which can provide a ready guide to all cases."[4] Instead, the *DuPont* court laid out thirteen factors that decision-makers should weigh when figuring out if consumers might confuse one mark with another.[5]

What are these thirteen factors? They cover everything from the obvious to the subtle. The first looks at "[t]he similarity or dissimilarity of the marks in their entireties as to appearance, sound, connotation and commercial impression."[6] The second examines "[s]imilarity or dissimilarity and nature of the goods or services."[7] The third considers whether the marks travel in "[t]he similarity or dissimilarity of established, likely-to-continue trade channels."[8] The fourth asks about "[t]he conditions under which and buyers to whom sales are made" which plays on impulse purchases versus careful, sophisticated buying decisions.[9] The fifth weighs "[t]he fame of the prior mark (sales, advertising, length of use)."[10] The sixth looks at "[t]he number and nature of similar marks in use on similar goods."[11] The seventh addresses "[t]he nature and extent of any actual confusion", or simply asking have people actually gotten confused?[12] The eighth flips that around: "[t]he length of time during and conditions under which there has been concurrent use *without* evidence of actual confusion."[13] The ninth examines "[t]he variety of goods on which a mark is or is not used (house mark, 'family' mark, product mark)."[14] The tenth considers "[t]he market interface between applicant and the owner of a prior mark," including any agreements or legal history.[15] The eleventh asks about "[t]he extent to which applicant has a right to exclude others from use of its mark on its goods."[16] The twelfth weighs "[t]he extent of potential confusion", is it minimal or substantial?[17] And the thirteenth serves as a catchall for "[a]ny other established fact probative of the effect of use" that might matter.[18]

B. The Promise of Systematic Analysis

[2] *In re* E.I. du Pont de Nemours & Co., 476 F.2d 1357, 1361, 177 U.S.P.Q. 563, 567 (C.C.P.A. 1973).
[3] *See* TMEP § 1207.01 (Nov. 2025 ed.); 4 J. THOMAS MCCARTHY, MCCARTHY ON TRADEMARKS AND UNFAIR COMPETITION § 24:30 (5th ed. 2024) (describing *DuPont* as the "leading case" applied in "thousands" of TTAB and Federal Circuit decisions).
[4] *DuPont*, 476 F.2d at 1361, 177 U.S.P.Q. at 567.
[5] *Id.* at 1361–63, 177 U.S.P.Q. at 567–69.
[6] *Id.* at 1361, 177 U.S.P.Q. at 567.
[7] *Id.*
[8] *Id.*
[9] *Id.*
[10] *Id.*
[11] *Id.*
[12] *Id.*
[13] *Id.*
[14] *Id.*
[15] *Id.*
[16] *Id.*
[17] *Id.*
[18] *Id.*

The court made clear these factors weren't meant to be a rigid formula, but rather a framework for thoughtful, comprehensive analysis.[19] Courts and examining attorneys should only consider "those relevant factors for which there is evidence in the record", recognizing that not every factor matters in every case.[20] The *DuPont* decision emphasized that "each case must be decided on its own facts" and that factors don't have to be weighed equally because "each may from case to case play a dominant role."[21] Here's where it gets interesting: while *DuPont* established that factors could play dominant roles, later courts clarified that "any one of the factors may control a particular case," meaning sometimes a single factor could be decisive.[22] This created tension in the framework. On one hand, courts should look at all relevant factors holistically. On the other hand, sometimes one factor might dominate or even decide the outcome.[23] The takeaway? The analysis "implies no mathematical precision, and a plaintiff need not show that all, or even most, of the factors listed are present in any particular case to be successful."[24]

C. Widespread Institutional Adoption

The *DuPont* framework quickly became the go-to standard for the Trademark Trial and Appeal Board ("TTAB"), which now applies these factors in every confusion case arising from opposition and cancellation proceedings.[25] When the Court of Customs and Patent Appeals was reorganized in 1982, its trademark work moved to the newly-created Federal Circuit, which inherited *DuPont* as binding precedent.[26] The Federal Circuit has been clear: when parties present evidence or argument relating to a specific DuPont factor, the TTAB must address that factor in its analysis rather than ignoring it.[27] Today, the USPTO uses the *DuPont* factors to evaluate hundreds of thousands of trademark applications each year, with examining attorneys relying on *DuPont's* framework to decide whether new marks would likely confuse consumers.[28]

While the *DuPont* factors technically apply only to USPTO proceedings and Federal Circuit appeals, their influence spreads far beyond. Nearly every federal circuit court adopted similar multifactor tests for trademark infringement cases.[29] For example, the Second Circuit uses the eight-factor *Polaroid* test.[30] Also, the Third Circuit draws from *Interpace* and *Scott Paper*[31] while the Ninth Circuit follows *Sleekcraft.*[32] Despite their different names and slight variations, these tests all share *DuPont*'s DNA: a non-exhaustive list of factors, holistic

[19] *Id.* at 1361, 177 U.S.P.Q. at 567.
[20] *Id.* at 1361–62, 177 U.S.P.Q. at 567–68.
[21] *Id.* at 1361, 177 U.S.P.Q. at 567.
[22] *In re* Dixie Rests., Inc., 105 F.3d 1405, 1406–07, 41 U.S.P.Q.2d 1531, 1533 (Fed. Cir. 1997).
[23] *Id.*
[24] *Wynn Oil Co. v. Thomas*, 839 F.2d 1183, 1186 (6th Cir. 1988).
[25] TMEP § 1207.01 (Nov. 2025 ed.).
[26] Federal Courts Improvement Act of 1982, Pub. L. No. 97-164, 96 Stat. 25 (establishing the Federal Circuit and transferring the CCPA's jurisdiction to it).
[27] *In re* Guild Mortg. Co., 912 F.3d 1376, 1380–82 (Fed. Cir. 2019) (vacating where TTAB failed to address *DuPont* Factor 8 despite record evidence).
[28] *USPTO Trademarks Dashboard*, U.S. PAT. & TRADEMARK OFF., https://www.uspto.gov/dashboard/trademarks/ (last visited Nov. 29, 2025) (reporting FY 2024 totals nearing 765,000 applications).
[29] BARTON BEEBE, *An Empirical Study of the Multifactor Tests for Trademark Infringement*, 94 CAL. L. REV. 1581, 1581–82 & app. A (2006).
[30] *Polaroid Corp. v. Polarad Elecs. Corp.*, 287 F.2d 492 (2d Cir. 1961).
[31] *Interpace Corp. v. Lapp, Inc.*, 721 F.2d 460 (3d Cir. 1983); *Scott Paper Co. v. Scott's Liquid Gold, Inc.*, 589 F.2d 1225 (3d Cir. 1978).
[32] *AMF Inc. v. Sleekcraft Boats*, 599 F.2d 341 (9th Cir. 1979).

weighing, no mechanical formulas, and no single dispositive factor.[33] The widespread adoption of these multifactor frameworks demonstrates how the legal profession has embraced this comprehensive approach to analyzing trademark confusion.[34]

II. The Judicial Rhetoric: All Factors Matter

A. The Mantra of Comprehensive Analysis

Even though courts acknowledge that some factors matter more than others, they consistently preach the gospel of comprehensive analysis. Courts continuously assert that *all* relevant *DuPont* factors must be considered.[35] The Federal Circuit keeps reminding the TTAB that it "must consider the DuPont factors about which there is evidence" and can't just ignore factors when parties have presented proof that bears on those factor's outcomes.[36] This comprehensive approach has become almost ritualistic, with courts describing the analysis as examining "the totality of the circumstances."[37] The Federal Circuit has shot down any attempts at shortcuts, noting that "there is no mechanical test" and "each case must be decided on its own facts."[38]

The TTAB starts virtually every confusion analysis with the same boilerplate language: "[w]hen determining likelihood of confusion, [the Board] must consider all of the probative evidence of record bearing on the likelihood of confusion."[39] This ritualistic opening signals the Board's commitment to comprehensive factor analysis, whether the answer is obvious or murky.[40] Courts justify this approach by arguing that it prevents arbitrary decisions and ensures all relevant evidence gets proper attention.[41] The idea is that systematically working through multiple factors protects against judicial mistakes and cognitive bias.[42]

This rhetoric goes beyond the Federal Circuit to every circuit court dealing with trademark confusion.[43] The Second Circuit describes its *Polaroid* factors as requiring "a flexible approach that avoids a wooden application" while still demanding attention to each relevant factor.[44] The Ninth Circuit says its *Sleekcraft* factors "are not exhaustive and other variables

[33] *See* RESTATEMENT (THIRD) OF UNFAIR COMPETITION § 21 cmt. a (AM. L. INST. 1995) (noting that multifactor tests avoid "mechanistic formula" and require holistic analysis); 4 MCCARTHY, *supra* note 3, § 23:19.50.

[34] 4 MCCARTHY, *supra* note 3, § 23:19.50 (explaining that nearly every federal circuit has adopted multifactor confusion tests that share the structure and spirit of the *DuPont* framework).

[35] *Guild Mortg.*, 912 F.3d at 1380–82 (vacating where the Board failed to address a *DuPont* factor supported by evidence).

[36] *Id.* at 1379.

[37] *Joseph Phelps Vineyards, LLC v. Fairmont Hldgs., LLC*, 857 F.3d 1323, 1329 (Fed. Cir. 2017) (describing the analysis as examining "the totality of the circumstances").

[38] *DuPont*, 476 F.2d at 1361.

[39] *In re* Reach Int'l, Inc., Serial No. 97/335,655, slip op. at 5 (T.T.A.B. July 2024) ("Our determination . . . is based on an analysis of all of the probative evidence of record bearing on a likelihood of confusion.") (citing *DuPont*); *In re* Majestic Distilling Co., 315 F.3d 1311, 1315 (Fed. Cir. 2003).

[40] *See* TMEP § 1207.01 (Nov. 2025 ed.) (describing standard TTAB practice of comprehensive factor analysis).

[41] RESTATEMENT (THIRD) OF UNFAIR COMPETITION § 21 cmt. a (arguing that systematic multifactor analysis prevents arbitrary decisions and ensures proper attention to all relevant evidence).

[42] CHRIS GUTHRIE, JEFFREY J. RACHLINSKI & ANDREW J. WISTRICH, *Blinking on the Bench: How Judges Decide Cases*, 93 CORNELL L. REV. 1 (2007) (finding that judges are susceptible to cognitive biases that structured analysis may help mitigate).

[43] *See* Beebe, *supra* note 29, at 1584–85 (documenting the rhetoric of comprehensive multifactor analysis across circuits).

[44] *Nabisco, Inc. v. Warner-Lambert Co.*, 220 F.3d 43, 46 (2d Cir. 2000) ("the evaluation of the *Polaroid* factors is not a mechanical process . . ."); *see also* Morningside Grp. Ltd. v. Morningside Cap. Grp., L.L.C., 182 F.3d 133, 142 (2d Cir. 1999) (warning against a "wooden application" of general rules within the *Polaroid* analysis).

may come into play depending on the particular circumstances."[45] Across jurisdictions, the message is consistent: comprehensive multifactor analysis is the gold standard.[46]

B. The Refusal to Establish Hierarchy

While courts sometimes admit that certain factors tend to carry more weight, they refuse to establish any formal pecking order among the *DuPont* factors.[47] The Federal Circuit has observed that similarity of marks and relatedness of goods are "often" the most important, but immediately adds the qualifier that "not all of the DuPont factors are relevant or of similar weight in every case."[48] This pattern of acknowledging reality while refusing to formalize it shows up repeatedly in confusion cases.[49]

The TTAB and Federal Circuit often call the first two *DuPont* factors "key considerations," but these observations never translate into official prioritization.[50] Instead, courts insist that even when marks are highly similar and goods are closely related, they must still examine other factors, especially when parties have introduced evidence about purchaser sophistication, actual confusion, or mark strength.[51] This insistence on comprehensive analysis continues even in cases where the outcome seems obvious from the first two factors alone.[52] The Federal Circuit has made clear that the Board makes a mistake when it fails to consider factors for which evidence exists, even if those factors probably won't change the result.[53]

Circuit courts show the same reluctance to establish clear hierarchies.[54] The Ninth Circuit has stated that some *Sleekcraft* factors "are much more important than others" and that "the relative importance of each individual factor will be case-specific," yet maintains that "no single factor is supposed to be dispositive."[55] The Second Circuit acknowledges that the trademarks themselves and the goods or services "typically are considered to carry the greatest weight," but warns against treating these as "solely determinative."[56] The Third Circuit says that “[t]he single most important factor in determining likelihood of confusion is mark similarity,” but emphasizes that even in cases of directly competing goods “the factor

[45] *Sleekcraft*, 599 F.2d at 348 ("[T]he factors are not exhaustive and other variables may come into play depending on the particular circumstances.").

[46] Beebe, *supra* note 29, at 1584–85 (observing that courts across jurisdictions describe comprehensive multifactor analysis as the gold standard).

[47] *Dixie*, 105 F.3d at 1407 (quoting *In re* E.I. du Pont de Nemours & Co., 476 F.2d 1357, 1361 (C.C.P.A. 1973), and citing *Opryland USA, Inc. v. Great Am. Music Show, Inc.*, 970 F.2d 847, 850 (Fed. Cir. 1992); *In re* Shell Oil Co., 992 F.2d 1204, 1206 (Fed. Cir. 1993)).

[48] *Dixie*, 105 F.3d at 1407; *DuPont*, 476 F.2d at 1361.

[49] *See* Beebe, *supra* note 29, at 1586–87 (noting the pattern of courts acknowledging practical realities while refusing to formalize them into doctrine).

[50] *In re* i.am.symbolic, LLC, 866 F.3d 1315, 1322 (Fed. Cir. 2017) ("The two key considerations are the similarities between the marks and the similarities between the goods.").

[51] *Guild Mortg.*, 912 F.3d at 1379–80.

[52] *Id.*

[53] *Id.* at 1380.

[54] Beebe, *supra* note 29, at 1587 (observing circuit courts' similar reluctance to establish formal factor hierarchies).

[55] *Multi Time Mach., Inc. v. Amazon.com, Inc.*, 804 F.3d 930, 935 (9th Cir. 2015) ("[S]ome of the *Sleekcraft* factors are much more important than others, and the relative importance of each individual factor will be case-specific.") (quoting *Brookfield Commc'ns, Inc. v. W. Coast Ent. Corp.*, 174 F.3d 1036, 1054 (9th Cir. 1999)).

[56] *Nabisco*, 220 F.3d at 46 ("the similarity of the marks and the proximity of the products . . . typically are considered to carry the greatest weight"); *Arrow Fastener Co. v. Stanley Wks.*, 59 F.3d 384, 400 (2d Cir. 1995).

regarding the similarity of marks may increase in importance, [yet] it does not eliminate the other factors entirely."[57]

This resistance to hierarchy creates an odd tension in confusion doctrine.[58] Courts acknowledge that certain factors matter more while insisting that formal hierarchy would be wrong.[59] The result is mixed messaging: litigants know from experience that mark similarity and goods relatedness drive outcomes, yet doctrine formally requires them to address all factors with apparently equal thoroughness.[60]

C. Circuit Variations (or Alleged Variations)

Different federal circuits have developed their own multifactor tests, raising questions about whether confusion analysis actually varies by jurisdiction.[61] The Second Circuit's eight-factor *Polaroid* test includes "defendant's good faith in adopting the mark," which *DuPont* doesn't explicitly list.[62] The Ninth Circuit's *Sleekcraft* factors include "defendant's intent in selecting the mark," reflecting concern about deliberate copying.[63] The Seventh Circuit uses a seven-factor test that explicitly considers "the degree of care likely to be exercised by consumers."[64] The Third Circuit applies a ten-factor test from *Interpace* that includes factors like "the likelihood the senior user will bridge the gap" and "other facts suggesting that the consuming public might expect the prior owner to manufacture a product in the defendant's market."[65] These differences suggest potential substantive variation in how circuits assess confusion.[66]

Yet despite these variations in wording, the core factors remain remarkably consistent across circuits.[67] Every circuit test includes mark similarity, goods relatedness, and most include purchaser sophistication, actual confusion, and mark strength.[68] The real question is whether differences in factor lists produce differences in outcomes, or whether the variations are just semantic.[69] If courts across circuits reach the same conclusions in similar cases despite different factor formulations, the apparent variation may be more rhetorical than real.[70] Without systematic empirical analysis comparing case outcomes across circuits while controlling for factual differences, the question remains largely speculative.[71] The rhetoric of circuit variation may mask an underlying uniformity in actual decisions.[72]

D. The Illusion of Flexibility

[57] *A&H Sportswear, Inc. v. Victoria's Secret Stores, Inc.*, 237 F.3d 198, 211 (3d Cir. 2000) ("[T]he single most important factor in determining likelihood of confusion is mark similarity.").
[58] Beebe, *supra* note 29, at 1588.
[59] *Id.*
[60] *Id.* at 1589.
[61] *See id.* at 1590–92 (comparing circuit tests).
[62] *Polaroid*, 287 F.2d at 495.
[63] *Sleekcraft*, 599 F.2d at 348–49.
[64] *Helene Curtis Indus., Inc. v. Church & Dwight Co.*, 560 F.2d 1325, 1330 (7th Cir. 1977).
[65] *Interpace Corp.*, 721 F.2d at 463.
[66] Beebe, *supra* note 29, at 1590 (noting that different factor formulations suggest potential substantive variation across circuits).
[67] *Id.* at 1591 (observing that despite wording variations, core factors remain remarkably consistent across circuits).
[68] *Id.*
[69] *Id.* at 1592 (questioning whether differences in factor lists produce differences in outcomes or are merely semantic).
[70] *Id.*
[71] *Id.* at 1593 (noting that without systematic empirical analysis, the question of circuit variation remains speculative).
[72] *Id.*

Courts consistently praise the multifactor framework's flexibility as one of its greatest strengths.[73] The Federal Circuit has described the *DuPont* factors as providing a flexible framework that can accommodate the wide variety of factual scenarios that arise in trademark disputes.[74] This flexibility supposedly allows courts to reach the "right" result in each case rather than being locked into rigid formulas.[75] Judges value being able to weigh factors differently depending on context and emphasize the considerations most relevant to the specific case before them.[76]

But flexibility without clear standards easily becomes unpredictability.[77] When no factor is necessarily dispositive and each must be weighed according to circumstances, parties struggle to assess their chances before investing serious money in litigation.[78] The same facts may lead different adjudicators to different conclusions depending on which factors they emphasize and how they weigh conflicting considerations.[79] What courts celebrate as flexible contextualization, litigants experience as outcome uncertainty.[80]

This unpredictability has real consequences for trademark owners and applicants trying to navigate the registration and enforcement system.[81] Sophisticated parties must prepare evidence and arguments on all thirteen *DuPont* factors because any might prove significant in a particular case.[82] The flexibility that allows courts to reach nuanced conclusions in unusual cases imposes substantial costs in typical cases where the outcome could be predicted from basic facts.[83] The framework creates a gap between what doctrine demands (comprehensive multifactor analysis) and what would serve litigants better (clear guidance about what actually matters).[84]

The celebration of flexibility also hides an important question: are courts genuinely exercising contextualized judgment, or are they following predictable patterns while maintaining the appearance of individualized analysis?[85] If confusion outcomes are actually highly predictable from a small number of factors, then the supposed flexibility may be largely illusory.[86] Courts may be deciding cases based on mark similarity and goods relatedness, then using other factors to construct after-the-fact justifications for conclusions

[73] *Dixie*, 105 F.3d at 1406–07 (explaining the flexibility of the *DuPont* framework).

[74] *In re* Shell Oil Co., 992 F.2d 1204, 1206 (Fed. Cir. 1993) ("[T]he various evidentiary factors may play more or less weighty roles in any particular determination."); *In re* Mighty Leaf Tea, 601 F.3d 1342, 1346 (Fed. Cir. 2010) ("Not all of the *DuPont* factors are relevant to every case.").

[75] *See* Beebe, *supra* note 29, at 1584 (noting courts' celebration of flexibility as allowing the "right" result in each case).

[76] *Id.* at 1584–85 (documenting judicial rhetoric emphasizing contextual weighing of factors).

[77] *Id.* at 1585 (observing that flexibility without clear standards produces unpredictability).

[78] *Id.*

[79] *Id.* at 1585–86 (noting that identical facts may yield different conclusions depending on which factors adjudicators emphasize).

[80] *Id.* at 1586.

[81] *See id.* at 1587 (discussing practical consequences of unpredictability for trademark owners and applicants).

[82] *See* TMEP § 1207.01(a) (Nov. 2025 ed.) (requiring consideration of all relevant *DuPont* factors); Beebe, *supra* note 29, at 1587–88 (noting litigants must prepare for all thirteen factors).

[83] Beebe, *supra* note 29, at 1588 (arguing that flexibility in unusual cases imposes costs in typical cases).

[84] *Id.*

[85] *Id.* at 1640–44 (questioning whether courts exercise genuine contextualized judgment or follow predictable patterns).

[86] *Id.*

already reached.[87] The rhetoric of comprehensive analysis would thus mask a simpler underlying decision rule.[88]

III. The Cost of the Framework

A. For Litigants: The Burden of Thirteen

The comprehensive nature of *DuPont* analysis hits litigants hard in the wallet.[89] The average trademark infringement lawsuit in the U.S. runs between $120,000 and $750,000, depending on complexity and whether it goes to trial.[90] For small businesses, costs typically range from $50,000 to $250,000 or more, depending on the case, jurisdiction, and legal representation.[91] These costs pile up because litigants are expected to develop evidence on all potentially relevant *DuPont* factors.[92]

Factor 7 (actual confusion) is particularly brutal.[93] Courts frequently describe properly designed consumer surveys as among the most probative evidence of actual confusion; while surveys are not strictly required, several courts have remarked that the absence of survey evidence can weigh against a party's case.[94] Because credible surveys require expert design, fielding, controls, and reporting, budgets often reach the **high five to low six figures**, and complex matters can exceed **$100,000**.[95] While courts consider well-designed surveys among the best evidence of confusion, the price tag often puts such evidence out of reach for smaller trademark owners.[96] Some courts have even stated that "the absence of surveys is evidence that actual confusion cannot be shown."[97] The result is a Catch-22: surveys are expensive to conduct, but their absence may be held against you.[98]

[87] *Id.* at 1641 (suggesting courts may decide based on a few key factors, then construct justifications from other factors).

[88] *Id.*

[89] AM. INTELL. PROP. L. ASS'N, 2023 REPORT OF THE ECONOMIC SURVEY (2023) (reporting median litigation costs for trademark cases at various stages).

[90] *See Average Cost of Trademark Infringement Lawsuit: Insights & Strategies*, ADIBI IP (May 30, 2025), https://adibiip.com/average-cost-of-trademark-infringement-lawsuit/ (last visited Nov. 30, 2025) (estimating costs between $120,000 and $750,000 depending on complexity).

[91] IGOR DEMČÁK, *Understanding the Financial Damages of Trademark Infringement: The Costly Consequences*, LEXOLOGY (July 19, 2023), https://www.lexology.com/library/detail.aspx?g=74cb5021-b17a-44c4-b25a-eb1dcd5b47c9 (last visited Nov. 30, 2025) (noting small business costs "range from $50,000 to $250,000 or more").

[92] *Guild Mortg.*, 912 F.3d at 1379 (holding it error to ignore *DuPont* factors supported by record evidence).

[93] Beebe, *supra* note 29, at 1605–07 (discussing the role, weight, and costs of actual-confusion evidence and surveys).

[94] *Sports Auth., Inc. v. Prime Hosp. Corp.*, 89 F.3d 955, 963–64 (2d Cir. 1996) ("[T]he absence of surveys is evidence that actual confusion cannot be shown.").

[95] ROBERT N. ENNS, *Practical Tips for Litigating Survey Evidence*, *in* ANNUAL MEETING COURSE MATERIALS 211, 213 (ABA Section of Intell. Prop. L. 2000) (noting survey costs "can vary widely, from a low of $15-20,000, to a high well into six figures").

[96] PETER HESS, GENNA LIU & HANA DAI, *What Recent Case Law Tells Us About the Importance of Consumer Surveys in Trademark Cases*, IPWATCHDOG (Aug. 31, 2021), https://ipwatchdog.com/2021/08/31/recent-case-law-tells-us-importance-consumer-surveys-trademark-cases/ (last visited Nov. 30, 2025) (discussing courts' treatment of well-designed surveys as reliable confusion evidence).

[97] *Sports Auth.*, 89 F.3d at 964; *see also Disney Enters., Inc. v. Sarelli*, 322 F. Supp. 3d 413, 442 (S.D.N.Y. 2018) (noting absence of survey evidence weighed against plaintiff).

[98] *See Sports Auth.*, 89 F.3d at 964; *cf. Majestic Distilling*, 315 F.3d at 1317 (noting in *ex parte* context, lack of actual-confusion evidence is often of "little evidentiary value" because applicant has no opportunity to gather such evidence).

Factor 4 (purchaser sophistication) often prompts parties to retain experts or present market evidence on consumer behavior and market conditions.[99] Factor 6 (number and nature of similar marks on similar goods) regularly entails assembling extensive third-party use and registration proof, evidence that is powerful but costly to gather and organize.[100] Factors 8 through 13, while often adding minimal value, still must be addressed because courts insist all relevant factors receive consideration.[101] The result? Hundreds of thousands of dollars per case, with discovery, expert fees, and briefing devoted to factors that rarely prove decisive.[102]

These costs hit small businesses and individual trademark owners disproportionately hard.[103] A telling U.S. example: in *CrossFit, Inc. v. Mustapha*, the court awarded $10,000 in statutory damages on the Lanham Act claim but $253,897.53 in fees and costs under Massachusetts Chapter 93A (more than 25× the damages), illustrating how litigation spend can dwarf monetary recovery.[104] Unless a business is prepared to invest hundreds of thousands or millions to protect a brand, sending a cease-and-desist letter and exploring settlement often becomes the more practical route, even with a strong case.[105] The complexity of thirteen-factor analysis thus prices small businesses out of enforcement, creating a real access-to-justice problem.[106]

B. For Courts: The Burden of Comprehensive Opinions

The institutional commitment to comprehensive *DuPont* analysis eats up substantial judicial resources.[107] TTAB decisions routinely span dozens of pages as the Board methodically addresses each relevant factor.[108] The Board issued more than 600 final decisions in 2022, with this output representing the culmination of lengthy analysis and deliberation.[109] Even when the outcome seems obvious from the first two factors (mark similarity and goods relatedness), the Board must still address other factors for which parties have introduced evidence.[110]

This ritualistic discussion continues even in cases where additional factors contribute nothing to the analysis.[111] The Federal Circuit has emphasized that the Board makes a mistake when it fails to consider factors for which record evidence exists, even if those factors seem

[99] TMEP § 1207.01(d)(vii) (Nov. 2025 ed.) (sophisticated purchasers less likely to be confused).
[100] *Juice Generation, Inc. v. GS Enters. LLC*, 794 F.3d 1334, 1338–39 (Fed. Cir. 2015); *Jack Wolfskin Ausrüstung für Draussen GmbH & Co. KGaA v. New Millennium Sports, S.L.U.*, 797 F.3d 1363, 1373–74 (Fed. Cir. 2015) (extensive third-party use/registrations can significantly weaken mark strength).
[101] *Guild Mortg.*, 912 F.3d at 1379–80.
[102] AM. INTELL. PROP. L. ASS'N, *supra* note 89 (reporting median trademark litigation costs)..
[103] Demčák, *supra* note 91.
[104] *CrossFit, Inc. v. Mustapha*, No. 13-11498-FDS, slip op. at 2, 5–6 (D. Mass. June 23, 2016) (order on attorneys' fees) (awarding $10,000 statutory damages under the Lanham Act and $229,881.08 in attorneys' fees plus $24,016.45 in costs under MASS. GEN. LAWS ch. 93A).
[105] JOSH GERBEN, *What to Expect in Trademark Litigation: A Step-by-Step Guide*, GERBEN IP (Apr. 28, 2025), https://www.gerbenlaw.com/blog/what-to-expect-in-trademark-litigation-a-step-by-step-guide/ (last visited Dec. 1, 2025).
[106] Beebe, *supra* note 29, at 1586.
[107] ROBERT G. BONE, *Taking the Confusion Out of "Likelihood of Confusion": Toward a More Sensible Approach to Trademark Infringement*, 106 NW. U. L. REV. 1307, 1309–10 (2012).
[108] *See, e.g.*, *Bd. of Trs. of Univ. of Ala. v. Pitts*, 107 U.S.P.Q.2d 2001 (T.T.A.B. 2013) (78-page precedential opinion); MARK A. JANIS & TIMOTHY R. HOLBROOK, *Patent Law's Audience*, 97 MINN. L. REV. 72, 111 (2012).
[109] JOHN L. WELCH, *Top 10 TTAB Decisions of 2022*, WORLD TRADEMARK REV. (Dec. 29, 2022), https://www.worldtrademarkreview.com/article/top-10-ttab-decisions-of-2022 (last visited Dec. 1, 2025).
[110] *Guild Mortg.*, 912 F.3d at 1379–80 (Board must consider each *DuPont* factor for which there is evidence).
[111] *Id.* at 1380.

unlikely to change the outcome.[112] In *In re Guild Mortgage Co.*, the Board's failure to address Factor 8 (concurrent use without confusion) required reversal and remand despite the factor's questionable significance.[113] The Court found that because the evidence weighed against confusion, the error couldn't be deemed harmless.[114] This creates pressure on the Board to address every factor exhaustively, lest an omission become grounds for reversal.[115]

Appellate review reinforces this comprehensive approach.[116] The Federal Circuit reviews TTAB decisions to confirm that the Board "considered" all relevant factors and properly weighed the evidence.[117] This standard encourages thorough discussion of each factor rather than focused analysis of the truly dispositive issues.[118] The time spent on a ritualistic factor analysis could be devoted to other cases or to more thoughtful consideration of genuinely difficult questions.[119] Instead, judicial resources are unnecessarily consumed by lengthy opinions that methodically work through factors that experienced practitioners know rarely matter.[120]

C. For the System: Unpredictability and Inconsistency

The flexibility that courts celebrate in the multifactor framework translates into unpredictability for litigants trying to assess their cases beforehand.[121] Scholars have criticized the likelihood of confusion test as producing "bad results," being "doctrinally incoherent," and lacking "a sensible normative foundation."[122] “The test is open-ended and subjective, producing uncertainty and expensive litigation."[123] When no single factor is dispositive and each must be weighed according to the totality of circumstances, parties struggle to predict outcomes before investing substantial resources in litigation.[124]

This unpredictability undermines settlement negotiations.[125] If both parties can't agree on the probable outcome, they can't easily agree on settlement value.[126] Each side may genuinely believe (based on different factors or different weighings) that it's likely to prevail.[127] The results are protracted disputes and unnecessary litigation costs.[128] Empirical studies suggest that outcomes are, in fact, quite predictable from the first two factors, but the doctrinal insistence on comprehensive analysis obscures this reality from the parties themselves.[129]

[112] *Id.* at 1379–81.
[113] *Id.* at 1381.
[114] *Id.* at 1381–82 (evidence on factor 8 weighed against finding of confusion).
[115] *See Guild Mortg.*, 912 F.3d at 1379–82; *Guild Mortg.*, 2020 U.S.P.Q.2d 10279, at *2–3 (T.T.A.B. 2020) (on remand, comprehensively addressing all factors).
[116] *Guild Mortg.*, 912 F.3d at 1379–80.
[117] *Id.* at 1380 ("the Board must consider each *DuPont* factor for which there is evidence").
[118] *See* Beebe, *supra* note 29, at 1588 (describing incentives created by comprehensive analysis requirements).
[119] *Id.* at 1640–44.
[120] *Id.* at 1585–86.
[121] *Id.* at 1586.
[122] Bone, *supra* note 107, at 1309–10.
[123] Note, *Trademark Injury in Law and Fact: A Standing Defense to Modern Infringement*, 135 HARV. L. REV. 667, 669 (2021).
[124] Beebe, *supra* note 29, at 1586–87.
[125] *See id.* at 1587.
[126] *See id.*
[127] *See id.*
[128] *Id.* at 1588.
[129] *Id.* at 1582–83.

Examining attorneys at the USPTO face similar challenges in achieving consistency across thousands of confusion determinations annually.[130] The *DuPont* framework says that "not all factors may be relevant" and that "any one of the factors may control a particular case," leaving substantial discretion to individual examiners.[131] Different examiners may weigh the same factors differently or emphasize different considerations.[132] The USPTO has implemented quality metrics and training programs to promote consistency, but the inherently flexible nature of the thirteen-factor test limits how much uniformity can be achieved.[133] The lack of clear guidance on factor hierarchy means examiner judgment plays a substantial role, introducing examiner-specific variation into what should be a more predictable administrative process.[134]

D. The Practiced Eye Knows Better

Experienced trademark practitioners have long recognized that mark similarity and goods relatedness drive outcomes in the vast majority of cases.[135] Barton Beebe's empirical study confirmed what sophisticated lawyers already knew from experience: judges use "fast and frugal" heuristics to short-circuit the multifactor analysis, with a few factors proving decisive while the rest are "at best redundant and at worst irrelevant."[136] Beebe found that "[a] finding that the similarity of the marks factor does not favor a likelihood of confusion is, in practice, dispositive, and a finding that the proximity of the goods factor does not favor a likelihood of confusion is nearly dispositive."[137]

Yet despite this practical understanding, litigants must still brief all thirteen *DuPont* factors to comply with doctrinal requirements.[138] Failure to address a factor for which evidence exists risks appellate reversal.[139] The result is ritual compliance: lawyers know which factors truly matter, but must pretend that all factors receive equal consideration.[140] This creates cognitive dissonance at the heart of trademark practice.[141] Attorneys counsel clients that cases with similar marks and overlapping goods are likely losers, but then must develop expensive evidence on purchaser sophistication, actual confusion, and other peripheral factors because doctrine demands comprehensive analysis.[142]

[130] U.S. PAT. & TRADEMARK OFF., *Trademarks Dashboard*, https://www.uspto.gov/dashboard/trademarks/ (last visited Dec. 1, 2025) (showing hundreds of thousands of trademark classes filed per year); U.S. PAT. & TRADEMARK OFF., *Likelihood of Confusion*, https://www.uspto.gov/trademarks/search/likelihood-confusion (last visited Dec. 1, 2025) ("[I]t's the most common reason for refusing registration.").

[131] *See* TMEP § 1207.01(a) (Nov. 2025 ed.) ("Not all of the *DuPont* factors may be relevant or of equal weight in a given case, and 'any one of the factors may control a particular case.'").

[132] *DuPont*, 476 F.2d at 1361 ("[T]here is no litmus rule which can provide a ready guide to all cases."); TMEP § 1207.01(a) (Nov. 2025 ed.).

[133] U.S. PAT. & TRADEMARK OFF., *Trademarks Dashboard*, https://www.uspto.gov/dashboard/trademarks/ (last visited Dec. 1, 2025) (quality metrics); U.S. PAT. & TRADEMARK OFF., TRADEMARK OPERATIONS UPDATE 8–10 (Apr. 28, 2023).

[134] Beebe, *supra* note 29, at 1584–85.

[135] *Id.* at 1581, 1582-83.

[136] *Id.* at 1640-1641 (describing judges' use of "fast and frugal" heuristics and noting that a few factors are decisive while others are "at best redundant and at worst irrelevant").

[137] *Id.* at 1582–83.

[138] *Guild Mortg.*, 912 F.3d at 1379–80.

[139] *Id.* at 1379–81.

[140] Beebe, *supra* note 29, at 1588.

[141] *Id.* at 1587–88.

[142] *Id.* at 1640–44.

The gap between what experienced practitioners know and what doctrine requires represents a profound inefficiency in the trademark system.[143] Resources get devoted to proving facts about factors that won't influence outcomes.[144] Courts write lengthy opinions addressing factors that don't drive their decisions.[145] Parties invest in litigation that predictability analysis (if honestly conducted) would show they're likely to lose.[146] The multifactor framework, celebrated for its flexibility and comprehensiveness, has become an expensive fiction that all participants maintain while privately knowing better.[147]

IV. Why the Framework Persists Despite Its Problems

A. Institutional Path Dependence

Fifty years of precedent built on the *DuPont* framework creates powerful institutional inertia that resists change.[148] The doctrine of *stare decisis* creates an explicitly path-dependent process where later decisions rely on and are constrained by earlier ones.[149] Once a judicial precedent has been established and relied upon, the costs of reversal grow through what scholars call "positive feedback."[150] In Hathaway's terms, the common law shows *increasing-returns path dependence*: each decision nudges future courts toward the same doctrinal path, raising the likelihood that subsequent decisions take a similar form.[151]

Circuit courts are bound by their own precedent under the law of the circuit doctrine, which requires three-judge panels to give *stare decisis* effect to past decisions that can only be overruled by the circuit sitting en banc or by the Supreme Court.[152] The TTAB is bound by Federal Circuit precedent, and when Congress created the Federal Circuit in 1982, the court adopted the CCPA's holdings as binding precedent.[153] In its first published opinion, the Federal Circuit adopted all CCPA (and Court of Claims) holdings issued before September 30, 1982, as binding precedent.[154] Because panels are bound by prior circuit precedent, abandoning *DuPont* would require en banc action by the Federal Circuit or Supreme Court review.[155]

In recent decades, the Supreme Court has rarely addressed the trademark likelihood-of-confusion framework on the merits.[156] While the Court granted certiorari in *B&B Hardware, Inc. v. Hargis Industries, Inc.* in 2014, that case addressed the preclusive effect of TTAB decisions rather than the substantive framework for assessing confusion.[157] Absent Supreme

[143] *Id.*

[144] *Id.*

[145] *Id.* at 1586.

[146] *Id.* at 1640–44.

[147] *Id.* at 1640–44.

[148] *DuPont*, 476 F.2d at 1357.

[149] OONA A. HATHAWAY, *Path Dependence in the Law: The Course and Pattern of Legal Change in a Common Law System*, 86 IOWA L. REV. 601 (2001).

[150] *Id.* at 605.

[151] *Id.* at 632 (explaining increasing-returns path dependence in law).

[152] JOSEPH W. MEAD, *Stare Decisis in the Inferior Courts of the United States*, 12 NEV. L.J. 787, 794–95 (2012) ("[E]ach circuit court has adopted some version of [the law of the circuit doctrine].").

[153] Pub. L. No. 97-164, 96 Stat. 25 (1982); *South Corp. v. United States*, 690 F.2d 1368, 1369–70 (Fed. Cir. 1982) (en banc).

[154] *South Corp.*, 690 F.2d at 1369–70.

[155] *See, e.g.*, FED. JUDICIAL CTR., THE ROLE OF THE U.S. COURTS OF APPEALS IN THE FEDERAL JUDICIARY, https://www.fjc.gov/history/work-courts/Role-of-the-Courts-of-Appeals (last visited Nov. 29, 2025); HENRY J. DICKMAN, *Note*, *Conflicts of Precedent*, 106 VA. L. REV. 1345, 1351–52 (2020).

[156] *See B&B Hardware, Inc. v. Hargis Indus., Inc.*, 575 U.S. 138 (2015).

[157] *Id.*

Court review, the law-of-the-circuit rule and the Federal Circuit's adoption of CCPA precedent keep the *du Pont* framework in place.[158] The accumulated precedent makes abandonment difficult without dramatic external pressure for change.[159]

B. Judicial Culture and Preferences

Judges value comprehensive analysis as a signal of thoroughness and careful consideration.[160] Discussing all relevant *DuPont* factors shows that the court has not overlooked potentially significant evidence.[161] Judicial opinions repeatedly stress flexibility and holistic weighing, emphasizing that there is no mechanical rule for likelihood of confusion.[162] The modern multifactor confusion test emerged as a compromise that "gave judges broad discretion to balance those factors as they saw fit."[163]

This preference for standards over hard-and-fast rules reflects deeper judicial values.[164] Scholars argue that bright-line rules can be inflexible, often ill-suited to accommodate case-specific nuance.[165] Supreme Court doctrine cautions that voluntariness cannot be resolved by any "infallible touchstone," and favors case-by-case assessments under a totality of the circumstances approach.[166] Standards, compared to rules, promote closeness of fit between legal doctrine and factual circumstances.[167] Judicial minimalism often favors standards over bright-line rules, enabling courts to avoid clear, sweeping resolutions in areas where incremental development is prudent.[168]

The multifactor framework embodies these judicial preferences.[169] Courts repeatedly emphasize that there is no mechanical rule and that each case turns on its own facts.[170] The Federal Circuit underscores flexibility: no mechanical rule determines likelihood of confusion, and not all *du Pont* factors are relevant in every case.[171] This flexibility allows courts to reach what they perceive as the "right" result in each case rather than being constrained by predetermined hierarchies.[172] Formalizing explicit primacy for two *du Pont* factors would depart from the Federal Circuit's repeated admonition that there is no mechanical rule and that only factors significant to the case need be considered.[173]

C. Strategic Ambiguity

[158] *South Corp.*, 690 F.2d at 1369–70; FED. JUDICIAL CTR., THE ROLE OF THE U.S. COURTS OF APPEALS IN THE FEDERAL JUDICIARY, https://www.fjc.gov/history/work-courts/Role-of-the-Courts-of-Appeals (last visited Nov. 29, 2025).
[159] Hathaway, *supra* note 149, at 640.
[160] *Guild Mortg.*, 912 F.3d at 1379.
[161] *Id.*
[162] *Shell Oil*, 992 F.2d at 1206 ("various evidentiary factors").
[163] GRAEME B. DINWOODIE & MARK D. JANIS, *Confusion Over Use: Contextualism in Trademark Law*, 92 IOWA L. REV. 1597, 1601–02 (2007).
[164] PIERRE SCHLAG, *Rules and Standards*, 33 UCLA L. REV. 379, 381–82 (1985).
[165] LOUIS KAPLOW, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557, 562–63 (1992).
[166] *Schneckloth v. Bustamonte*, 412 U.S. 218, 224–29 (1973); *Ohio v. Robinette*, 519 U.S. 33, 39 (1996).
[167] Schlag, *supra* note 164, at 382.
[168] CASS R. SUNSTEIN, ONE CASE AT A TIME: JUDICIAL MINIMALISM ON THE SUPREME COURT ix, 4–6 (1999); *cf.* MICHAEL COENEN, *Rules Against Rulification*, 124 YALE L.J. 576, 650–53 (2014).
[169] Dinwoodie & Janis, *supra* note 163, at 1601–02.
[170] *Shell Oil*, 992 F.2d at 1206; *DuPont*, 476 F.2d at 1361.
[171] *Shell Oil*, 992 F.2d at 1206; *In re* Mighty Leaf Tea, 601 F.3d 1342, 1346 (Fed. Cir. 2010).
[172] Dinwoodie & Janis, *supra* note 163, at 1601–02.
[173] *Shell Oil*, 992 F.2d at 1206; *Mighty Leaf Tea*, 601 F.3d at 1346.

Maintaining doctrinal flexibility serves strategic functions for courts.[174] When no factor is necessarily dispositive and all must be weighed holistically, judges retain maximum discretion to reach preferred outcomes.[175] Peripheral factors provide after-the-fact rationalization for decisions driven primarily by the first two factors.[176] Empirical studies suggest that judges use "fast and frugal" heuristics to reach conclusions based on mark similarity and goods relatedness, then "stampede" other factors to conform to the predetermined outcome.[177] The comprehensive framework obscures this reality and allows courts to maintain the appearance of individualized, thorough analysis while actually following predictable patterns.[178]

Flexibility also avoids committing to clear rules that might constrain future cases.[179] There is a well-recognized trade-off between commitment and flexibility: rules provide commitment ex ante but reduce flexibility ex post, while standards preserve discretion at the cost of certainty.[180] By refusing to establish explicit factor hierarchy, courts preserve discretion to emphasize different considerations in different contexts.[181] This strategic ambiguity allows outcomes to vary with changed circumstances or judicial preferences without requiring formal doctrinal revision.[182]

D. Lack of Systematic Evidence

Until the last two decades, large-scale empirical analysis of confusion cases had not been undertaken in any systematic way.[183] Barton Beebe's pioneering 2006 study examined 331 cases over a five-year period through labor-intensive hand-coding.[184] While groundbreaking, the study's limited scope prevented the drawing of definitive conclusions about temporal trends, circuit variations, or the predictive power of reduced models.[185] A 2009 follow-up study examining Southern District of New York cases over fifteen years confirmed Beebe's core findings but similarly lacked the scale to comprehensively challenge established doctrine.[186]

Hand-coding methodology inherently limits sample size.[187] Analyzing hundreds or thousands of cases requires months or years of work, making comprehensive temporal analysis or circuit comparisons impractical.[188] Academic incentives favor novel topics over replication, and legal scholars typically lack data science training while data scientists lack legal

[174] *See* Beebe, *supra* note 29, at 1640–44.
[175] *Id.*
[176] *Id*. at 1641.
[177] *Id*. at 1640.
[178] *Id*. at 1641.
[179] Schlag, *supra* note 164, at 386.
[180] Kaplow, *supra* note 165, at 562–66; LUCA ANDERLINI, LEONARDO FELLI & ALESSANDRO RIBONI, *Legal Efficiency and Consistency*, 121 EUR. ECON. REV. 103323 (2020).
[181] Beebe, *supra* note 29, at 1588; *Shell Oil*, 992 F.2d at 1206.
[182] Beebe, *supra* note 29, at 1640–44; *cf.* Sunstein, *supra* note 168, at 4–6.
[183] Beebe, *supra* note 29, at 1583–84.
[184] *Id.* at 1584–86.
[185] *Id.* at 1593–94 (discussing the limits of the dataset and the inferences that can be drawn).
[186] KEVIN BLUM ET AL., *Consistency of Confusion? A Fifteen-Year Revisiting of Barton Beebe's Empirical Analysis of Multifactor Tests for Trademark Infringement*, 2010 STAN. TECH. L. REV. 3.
[187] Beebe, *supra* note 29, at 1584–86, 1593–94.
[188] *See id.*; MARK A. HALL & RONALD F. WRIGHT, *Systematic Content Analysis of Judicial Opinions*, 96 CAL. L. REV. 63, 102 (2008) (noting that only a small minority of content-analysis projects code more than 1,000 cases).

expertise.[189] The result was a long delay between Beebe's 2006 study and the emergence of text-analysis tools capable of supporting truly large-scale empirical work on confusion cases.[190]

Without systematic proof that the thirteen-factor framework fails to function as advertised, courts had no empirical basis for abandoning fifty years of precedent.[191] Practitioners' intuitions and anecdotal observations, while suggestive, couldn't overcome institutional inertia absent comprehensive evidence.[192] The darkness persisted: experienced lawyers suspected that only two factors truly mattered, but lacked the data to prove it.[193] New tools in computational text analysis and artificial intelligence now make it feasible to run the kind of large-scale empirical studies of confusion doctrine that were simply not practical when Beebe wrote in 2006.[194]

[189] *See* MARINA KRAKOVSKY, *Just the Facts: Empirical Legal Studies on the Rise*, STAN. LAW., Spring 2009, at 24.

[190] *See* Beebe, *supra* note 29, at 1593–94; DARYL LIM, *Trademark Confusion Revealed: An Empirical Analysis*, 71 AM. U. L. REV. 1285, 1292–94 (2022) (noting that contemporary empirical work still relies on hand-coded samples in the low hundreds and discussing how algorithmic tools could enable larger-scale analyses).

[191] *See* Hathaway, *supra* note 149, at 640; Beebe, *supra* note 29, at 1583–84.

[192] Beebe, *supra* note 29, at 1584–85.

[193] *Id.* at 1602–03.

[194] *See* Lim, *supra* note 190, at 1290–94; MICHAEL A. LIVERMORE & DANIEL N. ROCKMORE, LAW AS DATA: COMPUTATION, TEXT, & THE FUTURE OF LEGAL ANALYSIS 3–6 (2019) (discussing how advances in computational methods enable large-scale empirical studies of legal doctrine).

PART 2:

DATA

I. Beebe's Breakthrough: The First Empirical Light

A. The 2006 Study That Changed the Conversation

1. Barton Beebe's Pioneering Work

Before we explore what has come since Beebe, it is appropriate to appreciate the impact of the work. Before 2006, the thirteen circuits' different multifactor tests for consumer confusion had played a central role in American trademark litigation, yet they'd received little academic attention and no empirical analysis.[195] Professor Barton Beebe's article, "An Empirical Study of the Multifactor Tests for Trademark Infringement," published in the California Law Review in 2006, changed that.[196] Beebe's study represented the first systematic empirical examination of how courts actually apply the *DuPont* factors and their circuit equivalents.[197] Rather than accepting judicial rhetoric about comprehensive multifactor analysis at face value, Beebe adopted a revolutionary approach: count what courts actually do, not what they say they do.[198]

The study examined all reported federal district court opinions for the five-year period from 2000 to 2004 where a multifactor test for confusion was used.[199] Working from an original dataset of 331 opinions, Beebe meticulously hand-coded each case to identify patterns in how judges applied the various factors.[200] This labor-intensive methodology required reading each opinion, extracting data on which factors were discussed, which party each factor favored, and the ultimate outcome.[201] The dataset and coding form were made publicly available in Excel format, demonstrating Beebe's commitment to transparency and replicability.[202]

2. The Core Methodology

Beebe's methodology involved identifying all reported federal district court opinions within his timeframe in which a multifactor likelihood-of-confusion test was substantially applied.[203] For each case, he coded several key variables: whether each of the thirteen factors was analyzed (yes/no), which party the factor favored (plaintiff or defendant), and the ultimate outcome (confusion found or not found).[204] This categorical coding approach (recording whether factors were discussed and their direction) allowed for statistical analysis of patterns across hundreds of cases.[205] Beebe acknowledged the limitations of hand-coding, noting that the labor-intensive nature of the process constrained sample size and prevented more granular analysis.[206]

The study presented the multifactor test as an ideal case study in legal multifactor decision-making and developed a methodology and theoretical toolkit for studying this form of legal analysis.[207] Drawing upon recent social science learning on cognition and decision-making,

[195] Beebe, *supra* note 29, at 1581.
[196] *Id.*
[197] *Id.* at 1582.
[198] *Id.* at 1583-84.
[199] *Id.* at 1584-86.
[200] *Id.*
[201] *Id.* at 1586–88.
[202] BARTON BEEBE, *An Empirical Study of the Multifactor Tests for Trademark Infringement: Supplementary Materials*, BARTONBEEBE.COM, https://bartonbeebe.com (last visited Dec. 1, 2025).
[203] Beebe, *supra* note 29, at 1584–86.
[204] *Id.* at 1586–88.
[205] *Id.*
[206] *Id.* at 1593–94.
[207] *Id.* at 1582.

Beebe brought empirical rigor to questions that had previously been addressed only through anecdote and intuition.[208]

B. Beebe's Key Findings

1. Factor 1 (Mark Similarity) Dominates
Beebe found that Factor 1 (similarity of the marks) was analyzed in nearly all of the 331 opinions in his dataset.[209] The factor showed strong correlation with outcomes, with confusion typically following when marks were found to be similar.[210] Beebe notes that courts have described the similarity factor as 'dispositive' and leading treatises call it 'usually controlling,' underscoring its centrality in the analysis.[211] Most strikingly, Beebe found that "a finding that the similarity of the marks factor does not favor a likelihood of confusion is, in practice, dispositive."[212] This finding suggested that despite judicial rhetoric about weighing all factors, mark dissimilarity effectively ends the inquiry.[213]

2. Factor 2 (Goods Similarity) Also Critical
Factor 2 (similarity and nature of the goods and services) was also analyzed in the vast majority of decisions, nearly as frequently as the similarity-of-the-marks factor.[214] Beebe found that "a finding that the proximity of the goods factor does not favor a likelihood of confusion is nearly dispositive."[215] The data revealed an interaction effect: cases involving both similar marks and similar goods were very likely to result in a finding of confusion, while dissimilarity on either dimension tended to be fatal to the plaintiff's case.[216]

3. Most Factors Rarely Analyzed or Determinative
Factors 6 through 13 were analyzed in only a minority of cases.[217] Factor 7 (actual confusion), while theoretically important, was often weakly developed in the record, with many opinions containing no survey or other direct evidence of confusion at all.[218] Factor 8 (concurrent use without confusion) was rarely present in the evidence.[219] Factors 9 through 13 received sporadic analysis and were usually found to be neutral.[220] When peripheral factors were analyzed, they appear rarely to change outcomes determined by the first two factors.[221]

Beebe discovered a counterintuitive finding regarding survey evidence: while many believe surveys to be the best and most persuasive form of evidence of confusion, the data revealed that surveys were rarely presented by parties or credited by courts.[222] Only 20% of the 331 opinions studied discussed survey evidence, and only 10% credited such evidence.[223]

4. The Illusion of Comprehensiveness

[208] *Id.* at 1583–84.
[209] *Id.* at 1605-06.
[210] *Id.*
[211] *Id.* at 1606-07.
[212] *Id.* at 1640.
[213] *See id.*
[214] *Id.* at 1608-09.
[215] *Id.* at 1640.
[216] *See id.* at 1608–09.
[217] *Id.* at 1610–20.
[218] *Id.* at 1615-17.
[219] *Id.* at 1616.
[220] *Id.* at 1617–20.
[221] *See id.* at 1640–44.
[222] *Id.* at 1615.
[223] *Id.*

Perhaps Beebe's most important finding was about judicial decision-making processes themselves.[224] Drawing on cognitive science research, Beebe showed that judges use "fast and frugal" heuristics to short-circuit the multifactor analysis.[225] A few factors prove decisive (primarily Factors 1 and 2) while the rest are "at best redundant and at worst irrelevant."[226] Crucially, judges tend to "stampede" these remaining factors to conform to the test outcome, particularly when they find infringement.[227] This stampeding effect means that once a judge decides based on mark similarity and goods relatedness, other factors are discussed in ways that support rather than test that initial conclusion.[228]

Courts claim to consider all factors comprehensively, but reality reveals overwhelming focus on Factors 1 and 2.[229] Other factors serve as window dressing and after-the-fact rationalization.[230] The thirteen-factor framework operates as ritual, not reality.[231]

C. The Limitations of Beebe's Study

1. Sample Size Constraints

Beebe's dataset of 331 cases over a five-year period (2000–2004), while groundbreaking for its time, imposed significant limitations.[232] The modest sample size necessarily constrained the statistical power for detecting subtle effects or temporal trends.[233] Five years of data are unlikely to capture how judicial practice may be evolving over longer time horizons.[234] The dataset was too small to support fully robust circuit-by-circuit analysis that could definitively answer whether rhetorical differences among circuits translated into outcome differences.[235] Labor-intensive hand-coding (Beebe coded all 331 opinions himself) prevented expansion to the larger sample that would have enabled more powerful statistical analysis.[236]

2. Methodological Constraints

Beebe's binary coding approach (recording whether each factor was analyzed and which party it favored) captured important patterns but missed potentially important nuances.[237] The methodology couldn't measure directional intensity: how strongly does a factor favor or disfavor confusion?[238] Without a scoring system, all "favors confusion" findings were treated equally, whether the Board found marks "virtually identical" or merely "somewhat similar."[239] This limitation made it difficult to model interaction effects between factors or to predict outcomes with precision.[240] A more granular scoring system (e.g., scoring Factor 1 from −5 to +5) might have revealed that Factor 1 scores of +5 overwhelm negative scores on other factors, but binary coding couldn't capture such relationships.[241]

[224] *Id.* at 1640.
[225] *Id.* at 1581, 1640.
[226] *Id.* at 1640.
[227] *Id.*
[228] *See id.* at 1640–44.
[229] *Id.* at 1641.
[230] *See id.*
[231] *Id.* at 1643.
[232] *Id.* at 1586.
[233] *Id.* at 1593–94.
[234] *See id.*
[235] *Id.*
[236] *Id.* at 1586.
[237] *Id.* at 1586–88.
[238] *See id.*
[239] *Id.* at 1588.
[240] *Id.* at 1593–94.
[241] *See id.*

3. Technology Limitations of the Era (2004)

Beebe conducted his study well before the advent of modern natural-language-processing pipelines and large language models; by contrast, recent work uses current NLP techniques to analyze millions of trademark records.[242] In 2000–2004, Beebe relied on hand-coding as the practical method for extracting structured data from legal opinions.[243] Although legal databases were fully digitized by 2000, Beebe did not have (or did not use) any computational pipeline for large-scale text analysis, relying instead on hand-coding.[244] The time-intensive nature of hand-coding (a task that, for a single researcher, realistically requires months of work) made larger-scale studies effectively impossible.[245] This technological constraint meant that even obviously valuable extensions of Beebe's work, such as analyzing 25 years of cases or thousands of decisions, remained beyond reach.[246]

4. Questions Left Open

Beebe's study proved that Factors 1 and 2 dominate, but left important questions unanswered.[247] Do Factors 3 through 5 (trade channels, purchaser sophistication, and mark strength) sometimes matter in genuinely close cases?[248] Has judicial practice changed in the two decades since Beebe's 2000-2004 sample period?[249] Can outcomes be predicted with high accuracy using just Factors 1 and 2, or do other factors occasionally swing close cases?[250] What about interaction effects between factors (does high mark similarity overcome weak goods relatedness in systematic ways)?[251] Do circuits genuinely differ in factor weighting, or is variation purely rhetorical?[252] These questions awaited technology that could analyze thousands of cases with the scoring granularity that Beebe's hand-coding couldn't achieve.[253]

D. The Twenty-Year Gap: What Happened After Beebe?

1. Follow-Up Studies

Several scholars undertook follow-up empirical work after Barton Beebe’s 2006 study, but none matched its breadth or altered doctrine in a significant way.[254] The most important early extension is Kevin Blum, Ariel Fox, Christina Hayes, and James Hanjun Xu’s article, *Consistency of Confusion? A Fifteen-Year Revisiting of Barton Beebe’s Empirical Analysis of Multifactor Tests for Trademark Infringement*.[255] Focusing on the Southern District of New York, Blum and his co-authors examined fifteen years of cases applying the Second Circuit’s

[242] SHIVAM ADARSH ET AL., *Automating Abercrombie: Machine-Learning Trademark Distinctiveness*, 21 J. EMPIRICAL LEGAL STUD. 826 (2024).

[243] Beebe, *supra* note 29, at 1586, 1643–55 (describing manual coding of the opinions).

[244] *Id.*

[245] *Id.* at 1586.

[246] *Id.*

[247] *Id.* at 1593–94.

[248] *Id.*

[249] *Id.*

[250] *Id.*

[251] *Id.* at 1640–44.

[252] *Id.* at 1590–92.

[253] *See id.* at 1593–94; *see also* Adarsh et al., *supra* note 242 (analyzing millions of USPTO records using machine-learning methods).

[254] Beebe, *supra* note 29.

[255] Blum et al., *supra* note 186.

Polaroid test, substantially expanding Beebe's 2000–2004 snapshot for that one influential district.[256]

Their results largely confirmed Beebe's core conclusions: only a small number of "key factors" actually drive outcomes, with similarity of the marks occupying the dominant position among them.[257] In that sense, Blum et al. provided an important replication and validation of Beebe's central claim that factor-weighting in confusion analysis is highly skewed, not evenly distributed across the multifactor framework.[258]

Beyond Blum et al., most subsequent empirical work remained narrow in scope. Some projects focused on specific circuits, particular industries, or limited time frames; others used content-analysis methods to study aspects of trademark doctrine adjacent to (but not identical with) the likelihood of confusion inquiry.[259] None, however, combined Beebe's national coverage, explicit focus on the multifactor test, and careful coding of factor-by-factor outcomes. His study remained the canonical empirical reference point for discussions of confusion analysis well into the 2020s.[260]

2. Why No Major Follow-Up?

Given the importance of Beebe's findings and the questions they left open, the absence of a large-scale, national follow-up study for nearly two decades requires explanation.[261] The primary constraint was methodological. Beebe's project depended on traditional content analysis: he and his research assistants identified all relevant opinions, read each one, and hand-coded whether each factor was discussed, which party it favored, and how the case ultimately came out.[262] Beebe emphasized that this "reading and coding each opinion" approach was extraordinarily labor-intensive. Scaling that method from 331 cases to several thousand would have required a prohibitive investment of time and research funding.[263]

More broadly, Beebe's project exemplified what Mark Hall and Ronald Wright later described as the promise and limits of systematic content analysis of judicial opinions.[264] Content analysis allows legal scholars to bring social-science rigor to questions about what courts actually do, but it also requires disciplined coding protocols, repeated reliability checks, and significant human effort.[265] The bottleneck is not the availability of opinions, as

[256] *Id.* (describing a dataset of Southern District of New York cases applying the Polaroid factors over a fifteen-year period and comparing results to Beebe's national study).
[257] *Id.* (reporting results "for the most part" consistent with Beebe's national study and emphasizing that similarity is paramount among the factors).
[258] *Id.* at 3–5 (using Beebe's 2000–2004 national dataset as a baseline and confirming that a small number of factors drive outcomes).
[259] *See, e.g.*, Lim, *Trademark Confusion Revealed*, *supra* note 190; DARYL LIM, *Trademark Confusion Simplified: A New Framework for Multifactor Tests*, 37 BERKELEY TECH. L.J. 867 (2022).
[260] BARTON BEEBE, *Empirical Studies of Trademark Law*, *in* 2 RESEARCH HANDBOOK ON THE ECONOMICS OF INTELLECTUAL PROPERTY LAW: ANALYTICAL METHODS 617 (Peter S. Menell & David L. Schwartz eds., 2019).
[261] *Id.* at 625–28 (noting the relative scarcity of large-scale empirical work on confusion after the 2006 study).
[262] Beebe, *supra* note 29, at 1586–88 (describing the process of reading and coding each opinion to record factor treatment and outcomes).
[263] *Id.* at 1586 (noting that the study required the author to read and code all 331 opinions in the dataset).
[264] Hall & Wright, *supra* note 188, at 63.
[265] *Id.* at 69–78 (describing the promises and limits of systematic content analysis, including the need for detailed coding schemes, reliability checks, and significant research time).

those are increasingly digitized, but the human capital necessary to transform unstructured judicial prose into coded data at scale.[266]

Institutional incentives compounded these methodological limits. Legal academia tends to reward novelty over replication. Extending Beebe's work would have meant many months (or years) of coding for a project that might be dismissed as "mere" replication rather than celebrated as a new theoretical contribution.[267] At the same time, the interdisciplinarity required for serious empirical work created a skills gap: many trademark scholars lacked training in statistics or data science, while empirically trained scholars often lacked the doctrinal knowledge and language familiarity necessary to code confusion factors reliably.[268] Marina Krakovsky's description of empirical legal studies as an emerging "third wave" in legal scholarship underscored both the promise and the difficulty of integrating sophisticated empirical methods into traditional doctrinal fields.[269]

Finally, technology simply had not advanced far enough to change the basic research production function. In the mid-2000s and 2010s, natural language processing tools remained too crude to assign nuanced, factor-by-factor labels across large corpora of judicial opinions without substantial human supervision.[270] Even as empirical legal studies flourished in other domains, trademark confusion remained, methodologically, where Beebe had left it: anyone who wanted more data had to be prepared to read and code more opinions by hand.[271]

3. The Waiting Period

In the years following publication, Beebe's 2006 article quickly became the definitive empirical account of how multifactor confusion tests operate in practice.[272] Scholars across doctrinal and theoretical camps cited his findings when critiquing the doctrinal incoherence of likelihood-of-confusion jurisprudence and the gap between multifactor rhetoric and actual decision-making.[273] The study's central claims that similarity of marks and proximity of goods dominate outcomes, and that other factors are often redundant or irrelevant became widely accepted in the academic literature.[274]

Yet academic consensus did not translate into doctrinal reform. Courts continued to recite and apply the full complement of *DuPont* or circuit-specific factors, insisting there was "no mechanical test" and that each case must be decided on its own facts, even as Beebe's data suggested that only a subset of factors did most of the work.[275] Practitioners, for their part, continued to brief all factors for which evidence existed, knowing that failure to address an "irrelevant" factor could invite criticism on appeal. Beebe's study sharpened critique and

[266] *Id.* at 75–76 (emphasizing that the primary constraint on sample size is typically the amount of trained human coding effort available).
[267] *See* Beebe, *Empirical Studies of Trademark Law*, *supra* note 260, at 625–30 (observing that empirical projects often require substantial investments of time and may not be rewarded as highly as doctrinal novelty).
[268] *See* Krakovsky, *supra* note 189, at 24, 26–27 (discussing the interdisciplinary expertise required for empirical legal studies and the challenges of collaboration between lawyers and empiricists).
[269] *Id.* at 24 (describing empirical legal studies as a "third wave" of legal scholarship following doctrinal and law-and-economics work).
[270] Adarsh et al., *supra* note 242, at 829–30 (contrasting traditional, hand-coded empirical projects with modern NLP- and ML-based approaches).
[271] Beebe, *supra* note 29, at 1586–88; Hall & Wright, *supra* note 188, at 75–76.
[272]*See* Beebe, *Empirical Studies of Trademark Law*, *supra* note 260, at 625–28.
[273]*See, e.g.*, Dinwoodie & Janis, *supra* note 163, at 1603–05; Lim, *Trademark Confusion Revealed*, *supra* note 190, at 1290–92.
[274] *See* Beebe, *supra* note 29, at 1582–83; Lim, *Trademark Confusion Revealed*, *supra* note 190, at 1290–92.
[275]*DuPont*, 476 F.2d at 1361; *Dixie*, 105 F.3d at 1406.

informed scholarly understanding, but it did not, by itself, provide the kind of overwhelming, longitudinal evidence that might justify doctrinal overhaul.

Some follow-up work further validated Beebe's conclusions. The Blum et al. study, examining fifteen years of Southern District of New York cases applying *Polaroid*, reported results "for the most part" consistent with Beebe's national dataset and again highlighted similarity of the marks as paramount.[276] More recently, Daryl Lim's empirical analysis of federal courts of appeals decisions found that judges frequently take "early off-ramps" in confusion cases by either "economizing" (analyzing only a handful of factors) or "folding" (collapsing multiple factors into one another).[277] Lim identified actual confusion, similarity of the marks, and competitive proximity as a "potent trio" that effectively guides the infringement inquiry in many cases.[278] In a subsequent article, he proposed a simplified framework to replace traditional multifactor tests, grounded in these empirical insights.[279] Despite these contributions, all prior empirical work on confusion shared a common structural limitation: manual or semi-manual coding of judicial opinions. Beebe's 331-case dataset required thousands of hours of human labor.[280] Blum and co-authors extended the time horizon to fifteen years, but only for a single district. Lim's work focused on courts of appeals, necessarily omitting the vast mass of TTAB proceedings and district-court opinions where confusion is applied most frequently.[281] The cognitive and temporal burden of reading, extracting, and coding factor discussions in thousands of multi-page decisions created a hard ceiling on sample size and complexity.[282]

That ceiling mattered most where the need for data was greatest. A truly comprehensive test of the multifactor framework would require a longitudinal, factor-by-factor analysis of hundreds or thousands of TTAB decisions, along with parallel treatment of district-court and appellate opinions. Until recently, such a project was effectively impossible without a small army of coders.[283]

Only in the 2020s did technology begin to catch up with the doctrinal questions Beebe had raised. Advances in machine learning and natural language processing, including transformer-based language models, made it feasible to automate large portions of the coding task that had constrained earlier projects.[284] Recent work in automated trademark analysis shows that models can classify distinctiveness, similarity, and even likelihood of confusion across hundreds or thousands of marks with high accuracy, using structured features extracted from text and, where available, images.[285] These methods do not eliminate the need

[276] Blum et al., *supra* note 186, at 3–5.
[277] Lim, *Trademark Confusion Revealed*, *supra* note 190, at 1347 (describing "early off-ramps" and distinguishing "economizing" and "folding" strategies).
[278] *Id.* at 1324–26 (identifying actual confusion, similarity of the marks, and competitive proximity as a "potent trio" that guides the infringement analysis).
[279] Lim, *Trademark Confusion Simplified*, *supra* note 259, at 869–72.
[280] Beebe, *supra* note 29, at 1586–88; Hall & Wright, *supra* note 188, at 75–76.
[281] Blum et al., *supra* note 186, at 3–5; Lim, *Trademark Confusion Revealed*, *supra* note 190, at 1292.
[282] Hall & Wright, *supra* note 188, at 75–76.
[283] *See id.*; Beebe, *supra* note 29, at 1586–88.
[284] Adarsh et al., *supra* note 242, at 829–36 (demonstrating modern NLP and ML techniques applied to large-scale trademark datasets).
[285] Adarsh et al., *supra* note 242; *see also* DARYL LIM, *Computational Trademark Infringement and Adjudication*, *in* RESEARCH HANDBOOK ON INTELLECTUAL PROPERTY AND ARTIFICIAL INTELLIGENCE 259, 262–66 (Ryan Abbott ed., 2022) (discussing how AI tools can assist in trademark dispute resolution).

for legal judgment, the need for human validation and careful prompt design remain, but they fundamentally change what is empirically possible.[286]

For nearly twenty years after Beebe, scholars and courts operated in a kind of empirical half-light: the best available evidence strongly suggested that only a few factors mattered, but no one had the tools or resources to prove it across the entire landscape of confusion decisions.[287] With modern natural language processing and large language models, that constraint has finally begun to lift. Comprehensive, scored, factor-by-factor analysis of thousands of TTAB and court decisions, previously beyond reach, has become a realistic project.[288]

Until now, the multifactor framework survived largely because no one could conclusively demonstrate, at scale, how little of it courts actually use. The next generation of empirical work has the potential to change that.[289]

II. The Technological Transformation: From Impossible to Inevitable

A. The LLM Revolution (2020-2024)

1. The Breakthrough Technologies

From 2020 into the mid-2020s, large language models stopped being curiosities and started looking like real research assistants.[290] OpenAI's GPT-3, released in 2020, was the first widely known model that could write reasonably coherent paragraphs, follow instructions, and answer questions across many domains from a single, general system.[291] It built on the transformer architecture introduced in 2017, which lets models track relationships across long passages rather than treating each word in isolation.[292] GPT-4, released in 2023, pushed that approach far enough to score at roughly the 90th percentile on a simulated Uniform Bar Exam and to perform strongly on other professional and graduate tests.[293] Newer systems like GPT-4.1 and GPT-5.1 added more stable long-form reasoning, better tool use, and "thinking" modes that let the model spend more time on harder problems, all features that make it more useful for complex legal research and exam-style questions.[294]

[286] Adarsh et al., *supra* note 242; Lim, *Computational Trademark Infringement and Adjudication*, *supra* note 285, at 262–66.

[287] *See* Beebe, *supra* note 29; Blum et al., *supra* note 186; Lim, *Trademark Confusion Revealed*, *supra* note 190.

[288] *See* Adarsh et al., *supra* note 242, at 829–36.

[289] *See* Lim, *Trademark Confusion Simplified*, *supra* note 259, at 869–72; Lim, *Computational Trademark Infringement and Adjudication*, *supra* note 285, at 262–66.

[290] JONATHAN H. CHOI, *How to Use Large Language Models for Empirical Legal Research*, 180 J. INSTITUTIONAL & THEORETICAL ECON. 214 (2024) (describing how LLMs enable new empirical legal research designs).

[291] Tom B. Brown et al., *Language Models Are Few-Shot Learners*, *in* ADVANCES IN NEURAL INFO. PROCESSING SYS. 33 (2020), https://arxiv.org/abs/2005.14165; *GPT-3*, WIKIPEDIA, https://en.wikipedia.org/wiki/GPT-3 (last visited Nov. 23, 2025).

[292] Ashish Vaswani et al., *Attention Is All You Need*, *in* 30 ADVANCES IN NEURAL INFO. PROCESSING SYS. 5998 (2017), https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[293] OpenAI, *GPT-4 Technical Report*, arXiv:2303.08774 (2023), https://arxiv.org/abs/2303.08774 (reporting GPT-4's performance on bar-exam-style and other professional tests); *GPT-4*, WIKIPEDIA, https://en.wikipedia.org/wiki/GPT-4 (last visited Nov. 23, 2025).

[294] *OpenAI, GPT-5.1: A Smarter, More Conversational ChatGPT*, OPENAI (Nov. 12, 2025), https://openai.com/index/gpt-5-1/; *OpenAI Reboots ChatGPT Experience with GPT-5.1 After Mixed Reviews of*

Anthropic's Claude family followed a slightly different path, with an early emphasis on safety, helpfulness, and long-document work. Claude 2 and 2.1 already posted impressive legal numbers: one early evaluation reported Claude 2 scoring around 76% on the multiple-choice portion of the bar exam, and Anthropic's own testing for Claude 3 Opus showed performance around 85% on Multistate Bar Examination–style questions and an LSAT score in the low 160s.[295] Claude 2.1 also introduced a 200,000-token context window, which is enough to read hundreds of pages at once.[296] The Claude 3 models refined that performance, and Claude Sonnet 4.5 is now pitched by Anthropic and legal-tech partners as a go-to model for litigation tasks: summarizing full briefing cycles, reviewing entire case records, and drafting first-cut judicial opinions.[297] Anthropic's own documentation recommends Sonnet 4.5 for legal summarization, and independent benchmarks rank the Claude 3 and Sonnet 4.x lines near the top on many legal and reasoning datasets.[298]

Google's Gemini line has taken yet another route, leaning heavily on integration with Google's broader ecosystem and very large context windows. Gemini 1.5 Pro was one of the first generally available models to offer million-token contexts, allowing it to take in millions of characters of text in a single call.[299] The later Gemini 3 Pro API maintains roughly a one-million-token input limit and currently sits at or near the top of public legal benchmark leaderboards (such as the Uniform Bar Exam section of the LLM-Stats "Legal" suite) alongside GPT-5-series and Claude-series models.[300] News coverage has also highlighted Gemini 3 Pro's performance on "Humanity's Last Exam," an ultra-difficult general-intelligence test that includes law among other domains, where it scored above other widely deployed models.[301] For law schools and legal-tech tools that live inside Google's world (Docs, Gmail, Drive), Gemini has quickly become a natural default.

GPT-5, VENTUREBEAT (Nov. 12, 2025), https://venturebeat.com/ai/openai-reboots-chatgpt-experience-with-gpt-5-1-after-mixed-reviews-of-gpt-5/.

[295] *Eden AI, Best Large Language Model APIs in 2023*, DEV.TO (Aug. 16, 2023), https://dev.to/edenai/best-large-language-model-apis-in-2023-24no (reporting that Claude 2 scored about 76.5% on the multiple-choice portion of the bar exam); *Claude AI Statistics and Insights 2025*, DATACAMP, https://www.datacamp.com/blog/claude-ai-stats (last visited Nov. 23, 2025) (summarizing Anthropic-reported standardized test performance for Claude 3 Opus, including LSAT and MBE-style scores).

[296] *Anthropic, Introducing Claude 2.1*, ANTHROPIC (Nov. 21, 2023), https://www.anthropic.com/news/claude-2-1 (describing a 200,000-token context window); *see also Long-Context Windows Get Huge*, *in* iS2 DIGITAL, AI: THE FUTURE OF CONTEXT: INDUSTRY TRENDS & EMERGING TECHNOLOGIES (2024).

[297] ANTHROPIC, THE CLAUDE 3 MODEL FAMILY: OPUS, SONNET, HAIKU (MODEL CARD) (2024), https://www.anthropic.com/claude; *Anthropic, Introducing Claude Sonnet 4.5*, ANTHROPIC (Sept. 29, 2025), https://www.anthropic.com/news/claude-sonnet-4-5 (quoting legal-tech partners on Sonnet's utility for legal research and document review).

[298] *Legal Summarization*, CLAUDE DOCS, https://platform.claude.com/docs/en/about-claude/use-case-guides/legal-summarization (last visited Nov. 23, 2025); *Claude AI Statistics and Insights 2025*, *supra* note 295 (reporting strong performance for Claude 3 models on legal and reasoning benchmarks).

[299] *Gemini 1.5 Pro and 1.5 Flash: Now With 2 Million Token Context Windows*, GOOGLE CLOUD BLOG (June 27, 2024), https://cloud.google.com/blog/products/ai-machine-learning/gemini-1-5-pro-generally-available (announcing 2-million-token contexts for Gemini 1.5 Pro); *What Is a Context Window?*, PINECONE, https://www.pinecone.io/learn/context-window/ (last visited Nov. 23, 2025).

[300] *Learn About Supported Models*, FIREBASE AI LOGIC, https://firebase.google.com/docs/ai/logic/models (last visited Nov. 23, 2025) (listing Gemini 3 Pro input token limit of 1,048,576 and output limit of 65,536); *LLM Benchmarks 2025 - Legal*, LLMSTATS, https://llmstats.com/benchmarks/legal (last visited Nov. 23, 2025) (showing Gemini 3 Pro at or near the top on Uniform Bar Exam–style benchmarks alongside Claude and GPT-series models).

[301] *New Google AI Posts Top Marks in "Humanity's Last Exam"*, AOL (Nov. 2025), https://www.aol.com/articles/google-ai-posts-top-marks-115919449.html (reporting Gemini 3 Pro's top score on a difficult general-intelligence exam) (reporting Gemini 3 Pro's top score on a difficult general-intelligence exam).

Across all three families, one technical change matters more than any branding: how much text the model can "see" at once. Early GPT-3 models were trained with a context window of about 2,048 tokens, which would amount to just a few pages of an opinion.[302] GPT-4 and GPT-4-Turbo expanded that to tens of thousands and then roughly 128,000 tokens; Claude 2.1 raised the ceiling to about 200,000 tokens; and newer flagships like GPT-4.1, Claude Sonnet 4/4.5, Gemini 1.5 Pro, and Gemini 3 Pro support context windows on the order of one million tokens.[303] In rough terms, that is enough for several casebooks' worth of text. For empirical trademark work, this means that instead of hand-coding one DuPont decision at a time, researchers can load entire TTAB opinions, even batches of them, into a single prompt and ask the model to identify which factors were discussed, which side each factor favored, and how strongly.[304] Recent work in empirical legal studies shows that, with careful prompts and auditing, these models can code legal texts at scale with accuracy comparable to trained research assistants, but at a speed and volume that were simply impossible a decade ago.[305]

2. What Changed for Legal Research

Large language models changed empirical legal research in a very specific way. They did not simply make things "faster" in some vague sense. They changed the tradeoff between scale and detail in a way that finally made comprehensive studies of confusion doctrine realistic.[306]

Earlier empirical work always had to choose between breadth and granularity. Barton Beebe's pioneering 2006 study analyzed 331 federal district court opinions over a five-year period.[307] Kevin Blum and co-authors later revisited his methodology over fifteen years of Southern District of New York cases applying the *Polaroid* test, but only for that single district.[308] Daryl Lim's 2022 study focused on federal courts of appeals, leaving out the thousands of TTAB proceedings where the *DuPont* factors are actually applied most often.[309] Every one of these projects required thousands of hours of human reading, extraction, and coding. In practice, that meant that a comprehensive, factor-by-factor analysis of TTAB decisions was out of reach. LLMs changed that cost structure.

Five capabilities mattered most: scale, granularity, consistency, reproducibility, and validation.

a. Scale

[302] Brown et al., *supra* note 291; *GPT-3*, *supra* note 291.

[303] *GPT-4 Technical Report*, *supra* note 293; *What Is a Context Window?*, *supra* note 299 (discussing GPT-3.5 and GPT-4 context limits); *OpenAI Debuts Its GPT-4.1 Flagship AI Model*, THE VERGE (Apr. 14, 2025), https://www.theverge.com/news/647896/openai-chatgpt-gpt-4-1-mini-nano-launch-availability (reporting a one-million-token context window for GPT-4.1); *Context Windows*, CLAUDE DOCS, https://platform.claude.com/docs/en/build-with-claude/context-windows (last visited Nov. 23, 2025) (noting one-million-token context for Claude Sonnet 4 and 4.5); *Learn About Supported Models*, *supra* note 300 (listing Gemini 3 Pro's token limits).

[304] *See Legal Summarization*, *supra* note 298 (describing use of Claude models to summarize long legal documents); Adnan Masood, *Long-Context Windows in Large Language Models: Applications in Comprehension and Code*, MEDIUM (Apr. 25, 2025), https://medium.com/@adnanmasood/long-context-windows-in-large-language-models-applications-in-comprehension-and-code-03bf4027066f (discussing how million-token context windows enable document-scale analysis).

[305] Choi, *supra* note 290, at 215–22; Adarsh et al., *supra* note 242.

[306] Choi, *supra* note 290.

[307] Beebe, *supra* note 29, at 1584.

[308] Blum et al., *supra* note 186, at 4–5.

[309] Lim, *Trademark Confusion Revealed*, *supra* note 190, at 1292.

LLMs can read and classify legal documents at speeds that would be impossible for human coders.[310] Jonathan Choi's study of Supreme Court opinions shows the basic pattern. In a simple classification task, GPT-4 matched the accuracy of trained research assistants but produced labels orders of magnitude faster.[311] What had once taken months of hand-coding can now be done in days.

The same logic applies to confusion decisions. Hand-coding 331 cases for Beebe's study required a substantial investment of time and research support.[312] With LLM-based extraction, coding several thousand TTAB and court opinions for factor presence and direction is realistic within a single project cycle. Instead of choosing between "a detailed study of a few hundred cases" and "a more superficial study of many cases," LLM-based workflows allow detailed coding at scale.

b. Granularity

Traditional automated text analysis was usually limited to keyword searches or very simple classifications. LLMs can extract much more nuanced information from legal text.[313] In a recent study of attribute extraction from legal documents, Adhikary and co-authors used large language models to pull structured information from judgments and demonstrated that LLMs could reliably map complex doctrinal language into detailed, labeled fields.[314]

For confusion analysis, that means a model can be asked not just whether an opinion discusses Factor 1, but how the tribunal describes the similarity of the marks. A model can distinguish "marks are virtually identical" from "marks share some elements but differ in overall commercial impression," and record where on a scale the decision falls. That allows intensity coding on, for example, a −5 to +5 scale rather than a simple yes/no.[315] LLMs can also return supporting quotations for each score, which makes it easy to check whether the extracted data matches the actual language of the opinion.[316]

c. Consistency

Human coders get tired, change their minds about borderline cases, and sometimes interpret coding guidelines differently from one another. LLMs are not "perfectly consistent," but they can be made more consistent than large teams of human coders when given clear, stable prompts.[317]

Li Wang and co-authors study this problem directly. They show that prompt engineering, combined with explicit scoring rubrics and examples, can significantly improve the

[310] *See* Choi, *supra* note 290, at 215–22.

[311] *Id.* at 216–19 (reporting that GPT-4 performed approximately as well as human coders in a Supreme Court classification task while operating much more quickly).

[312] *See* Beebe, *supra* note 29, at 1586–88.

[313] *See generally* Qiao Jin et al., *Demystifying Large Language Models for Medicine: A Primer*, arXiv:2410.18856 (2024) (discussing how LLMs can extract structured information from complex professional text); Choi, *supra* note 290, at 215–22.

[314] Subinay Adhikary, Procheta Sen, Dwaipayan Roy & Kripabandhu Ghosh, *A Case Study for Automated Attribute Extraction from Legal Documents Using Large Language Models*, ARTIF. INTELL. & L. (2024), https://doi.org/10.1007/s10506-024-09425-7.

[315] *See id.*; Choi, *supra* note 290, at 220–22.

[316] Adhikary et al., *supra* note 314; Choi, *supra* note 290, at 219–21.

[317] *See* Choi, *supra* note 290, at 216–22 (showing relatively stable performance from LLMs under fixed prompts).

consistency and reliability of LLM outputs across repeated runs and different models.[318] In practice, that means a researcher can write a detailed prompt that defines each *DuPont* factor, describes what counts as "strongly favors confusion" versus "slightly favors confusion," and then apply that same prompt to every case in the dataset. The model will still make mistakes, but those mistakes will at least be made within a fixed, documented framework.

d. Reproducibility

Hand-coded projects depend heavily on human judgment that is hard to describe in full. Even when authors publish their coding forms, much of the nuance lives in unwritten habits and one-off decisions. LLM-based workflows can be more reproducible because the core of the method is text.[319]

Choi's empirical study illustrates this advantage. He publishes the full prompts he used to instruct GPT-4, explains the classification tasks in detail, and compares model outputs to ground-truth labels.[320] Any later researcher can reuse those prompts, adjust them, or apply them to new corpora. LLM-based studies of confusion can do the same: share the entire prompt that defines each factor, provide examples of correct and incorrect outputs, and specify how the model's scores were converted into numeric variables. That kind of transparency is difficult to achieve when dozens of human coders are making thousands of small, undocumented decisions.

e. Validation

Finally, LLM workflows make systematic validation easier, not harder. In Choi's study, model classifications are compared directly to human labels on a held-out set of Supreme Court opinions, which allows a straightforward estimate of accuracy.[321] Similar strategies can be used for confusion cases. Researchers can spot-check random samples, compare model scores to human coders on a subset of opinions, and compute accuracy or agreement measures.[322]

Attribute extraction work in legal contexts follows the same pattern. Adhikary and colleagues evaluate LLM outputs against manually constructed ground truth and report performance on standard metrics such as precision and recall.[323] In addition, the structure of LLM outputs allows for internal checks. If a model says that Factor 1 "strongly favors confusion" and assigns a numeric score of +1, that mismatch can be flagged automatically for review. Extracted quotations can be used to confirm that the model has correctly summarized what the tribunal actually said. Outliers, such as opinions where the model reports an implausible combination of factors, can be triaged for manual inspection.[324] Together, these checks provide a level of documented quality control that is rarely feasible in purely manual projects.

B. Methodology of the Instant Study

1. From Beebe's 331 to 4,000 Cases

[318] Li Wang et al., *Prompt Engineering in Consistency and Reliability with the Evidence-Based Guideline for LLMs*, 7 NPJ DIGIT. MED. 41 (2024) (showing that carefully designed prompts improve consistency and reliability across models and tasks).

[319] *See id.*; Choi, *supra* note 290, at 221–22.

[320] Choi, *supra* note 290, at 215–22.

[321] *Id.* at 216–19 (comparing GPT-4 outputs to human-coded ground truth on Supreme Court opinions).

[322] *See id.* at 216–22 (describing accuracy checks and error analysis); Wang et al., *supra* note 318.

[323] Adhikary et al., *supra* note 314 (evaluating LLM outputs against manually constructed labels using precision and recall).

[324] *See* Choi, *supra* note 290, at 219–22; Adhikary et al., *supra* note 314; Wang et al., *supra* note 318 (describing techniques for detecting inconsistencies and improving reliability).

Beebe's 2006 article was a landmark, but it was also a product of its time. He read and hand-coded 331 federal district court opinions over roughly four years of decisions, a dataset supported powerful insights, yet it was inherently limited by the number of cases one person (plus research assistants) can realistically code.[325] No one volunteers to hand-code a few thousand opinions for fun.

This study takes a different approach. It analyzes roughly 4,000 TTAB decisions over a twenty-five-year period using an LLM-based pipeline built around Claude Sonnet 4.5. That is about a twelve-fold increase in sample size over Beebe's dataset and about six times the temporal span. Instead of simple "factor discussed / factor not discussed" variables, each confusion comparison in each opinion received: A numerical score from −5 to +5 for all thirteen *DuPont* factors; Directional labels (favors registrant, favors opposer, neutral); Verbatim supporting quotations from the opinion; Flags for special patterns, such as alternative grounds for decision or unusual factor combinations; and Metadata that supports time-series analysis, circuit and panel comparisons, and other statistical tests. In other words, the dataset is not only much larger than Beebe's. It is also much richer at the level of individual decisions.

Beebe himself explained why he stopped at 331 cases. Expanding the sample would have required prohibitively more time and money.[326] Binary coding was all that was feasible. A factor either was analyzed or was not; it either favored the plaintiff or the defendant.[327] With that sample size, there was no realistic way to test circuit-by-circuit differences or subtle temporal trends. Those limitations did not reflect any failure of Beebe's method. They reflected the simple fact that human beings can only read and code so fast.

Large language models remove a large part of that constraint. The present study uses Claude Sonnet 4.5 as the primary engine for data extraction. Anthropic's own documentation and partner testimonials pitch Sonnet 4.5 as state of the art for complex legal tasks, including analyzing full briefing cycles and producing draft judicial opinions, and as an excellent choice for high-accuracy legal summarization.[328] Claude's technical documentation also confirms that Sonnet 4 and 4.5 support context windows large enough to comfortably hold a full TTAB opinion, so no opinion needs to be split across multiple prompts.[329] That stability and capacity make Sonnet 4.5 a natural fit for opinion-level coding rather than chatty back-and-forth.

The pipeline worked at the level of a single confusion comparison at a time. Each TTAB opinion was sent to the model individually, with a highly detailed prompt that; Defined each of the thirteen *DuPont* factors; Explained what scores from −5 to +5 should mean for each factor; Instructed the model to identify every distinct mark-to-mark and class-to-class comparison that the Board actually analyzed; and Required the model to return, for each

[325] Beebe, *supra* note 29, at 1584, 1586 (describing a dataset of 331 reported federal district court opinions from 2000 through 2004 and noting the hand-coding required).
[326] *Id.* at 1593–94 (discussing the limits of expanding the dataset given time and resource constraints).
[327] *Id.* at 1586–88 (explaining the binary coding method used for factor presence and direction).
[328] *Claude Sonnet 4.5*, ANTHROPIC, https://www.anthropic.com/claude/sonnet (last visited Nov. 23, 2025) (quoting CoCounsel's description of Sonnet 4.5 as "state of the art on the most complex litigation tasks, for example, analyzing full briefing cycles and conducting research to synthesize and contrast arguments across documents").
[329] *Context Windows*, *supra* note 303; *see also Claude 4.5 Context Length & Extended Memory Explained*, SKYWORK AI (Oct. 2025), https://skywork.ai/blog/claude-4-5-context-length-extended-memory/ (explaining that Sonnet 4 and 4.5 can process up to one million tokens for eligible users, easily covering even the longest TTAB decisions).

factor in each comparison, both a numerical score and a short supporting quotation from the opinion. To reduce randomness, the model was called via API with a low temperature (0.2). The goal was not creativity. The goal was to behave like a very fast, very literal research assistant that never gets bored with long TTAB opinions.[330]

Because the prompt asked the model to parse comparisons explicitly, one TTAB opinion could yield multiple confusion analyses. If the Board compared one applicant's mark to two registrants' marks, or analyzed confusion across several classes of goods, the model treated each of those as a separate "case" for purposes of the dataset. This mirrors how practitioners and the Board think about confusion. The opinion is the container. The comparisons are where the action happens.

The system also took advantage of the model's ability to notice oddities. The prompt instructed Sonnet 4.5 to flag any opinion that seemed unusual. These flagged cases were pulled into a review queue for human inspection. In practice, the model did surface genuinely interesting outliers which will be the subject of later case studies.

The quote extraction step supported a second layer of tooling. For each opinion, the system generated a report that displayed the opinion text with the model's factor scores in the margins. Quotations that justified each score were highlighted next to the relevant paragraphs. A human reviewer could scroll the opinion, see that the model assigned a +4 to Factor 1 for a particular comparison, and immediately check the exact language that supposedly justified that score. This design aligns with recent work on LLM based attribute extraction in legal texts, which emphasizes rationales and traceability as key safeguards.[331] It also dramatically reduces the cost of human verification, since reviewers can focus on highlighted blocks instead of rereading entire opinions.

From a distance, the approach looks very different from Beebe's. At a conceptual level, however, it is an extension of his basic insight. Beebe showed that it was possible to take confusion opinions seriously as data and to code the factors in a structured way.[332] The present study uses a different tool and a bigger canvas. With a model like Sonnet 4.5 doing the first pass, it becomes feasible to extend Beebe's logic from 331 hand-coded cases to thousands of TTAB decisions while adding much more detailed information about how each factor was applied.[333]

2. Addressing Common Methodological Objections

Any empirical project that leans heavily on large language models invites skepticism. That skepticism is healthy. This Section briefly addresses the most common concerns about LLM-based legal coding and explains why, in this setting, they are important but not fatal.

a. "LLMs hallucinate"

LLMs sometimes produce confident but false statements. That problem is now well documented in legal contexts. Dahl and coauthors find legal hallucination rates between roughly 58 percent and 88 percent for general-purpose chatbots asked specific questions

[330] *See* Choi, *supra* note 290, at 219–22 (recommending low temperature settings for classification tasks to reduce stochastic variation and improve reproducibility, and describing LLMs as potential replacements for human research assistants in coding judicial opinions).

[331] *See generally* Adhikary et al., *supra* note 314.

[332] Beebe, *supra* note 29, at 1582–83.

[333] Choi, *supra* note 290, at 216–22.

about random federal cases.[334] Magesh and coauthors later show that even RAG-based legal research tools that market themselves as "hallucination free" still hallucinate between about 17 percent and 33 percent of the time on carefully designed benchmark queries.[335]

So the concern is real. It is also not unique. Human coders make mistakes too, especially when they are tired, rushed, or facing ambiguous text. Choi's study of Supreme Court opinions, for example, compares GPT-4 to trained research assistants and finds that GPT-4 performs approximately as well as the humans on a simple classification task, while being considerably faster.[336]

The relevant question, therefore, is not whether LLMs are perfect. They are not. The question is whether, for the specific task of extracting factor scores and quotations from TTAB opinions, they achieve accuracy comparable to human coders and whether their errors can be detected and corrected. The design choices in this project are aimed precisely at that goal: the model is given a detailed scoring rubric, run at low temperature, asked to return supporting quotations, and audited through spot checks and flagged outliers. In other words, the model is treated the way one would treat a junior research assistant who is fast but occasionally overconfident.

b. "Results depend on the prompt"

LLM outputs do depend on the prompt. That is a feature, not a hidden bug. Human coding is also judgment-dependent. Different coders, or the same coder on different days, may interpret an opinion differently. The difference is that human judgment usually lives in training sessions, email chains, and half-remembered conversations. LLM judgment lives in text.

Choi's methodological paper treats prompts as part of the research design and publishes them alongside results.[337] Wang and coauthors show that careful prompt engineering, including explicit role definitions and scoring rubrics, can improve the consistency and reliability of LLM outputs across tasks and models.[338] In this study, the full prompts that define each DuPont factor, each point on the −5 to +5 scale, and the coding rules for ambiguous cases can be reproduced in an appendix or online repository. That makes the judgment calls visible and contestable in a way traditional hand-coding rarely is.

c. "It is all a black box"

The internal workings of GPT-series, Claude-series, and Gemini-series models are complicated. That much is true. For empirical purposes, however, what matters is not whether we can describe every weight in a transformer, but whether the application of the model is transparent.

Here, the application looks more like a structured protocol than a mysterious oracle. The prompts are fixed. The temperature is set low. Each opinion is processed independently. The

[334] Matthew Dahl, Varun Magesh, Mirac Suzgun & Daniel E. Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, 16 J. LEGAL ANALYSIS 64, 69–70 (2024) (finding that LLMs hallucinate between 58% (ChatGPT 4) and 88% (Llama 2) of the time when asked direct, verifiable questions about random federal court cases).

[335] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, 22 J. EMPIRICAL LEGAL STUD. 1, 3–5 (2025) (demonstrating that RAG-based tools from LexisNexis and Thomson Reuters hallucinate between 17% and 33% of the time despite marketing claims).

[336] Choi, *supra* note 290, at 216–19.

[337] *Id.* at 215–22 (reproducing prompts and coding instructions used in the empirical study).

[338] Wang et al., *supra* note 318, at 1–8 (showing that carefully designed prompts, including explicit role definitions, improve consistency and reliability across models and tasks).

model returns explicit scores, short explanations, and quotations that support each score. Those outputs are then used to produce annotated opinions where human reviewers can see, in context, why the model thought Factor 1 should be +4 rather than +2.

Compared to a team of human coders, this setup is arguably less of a black box. Human coders bring years of tacit expertise and idiosyncratic habits that are difficult to document fully. An LLM, by contrast, will follow the same written instructions in exactly the same way every time, until the prompt or settings change. The "black box" critique is accurate about the underlying model architecture. It is less accurate about a research workflow designed to be documented and auditable from end to end.[339]

d. "LLMs do not really understand law"
This objection raises a deep philosophical question and, for purposes of this Article, an unhelpful one. The present project does not ask the model to decide hard questions of first impression or to generate binding doctrine. It asks the model to read TTAB opinions and answer much more mundane questions: Which DuPont factors did the Board discuss? Did the Board characterize Factor 1 as strongly favoring or strongly disfavoring confusion? What language did the Board use to explain that characterization?

Choi's empirical work suggests that LLMs can perform this sort of extraction task about as well as trained research assistants when given clear instructions and evaluated against ground-truth labels.[340] Adhikary and colleagues show that LLMs can map complex judicial text into structured attributes with high measured accuracy when evaluated against manually coded data.[341] Whether this counts as "genuine legal understanding" is a question for philosophers. For present purposes, it is enough that the model can label opinions in a way that aligns with how human coders would have labeled them, and that its rationales can be checked.

e. Scale versus perfection
Finally, there is a simple tradeoff. One option is to aim for near-perfect manual coding of a few hundred cases, as Beebe did. The other option is to accept slightly noisy LLM-assisted coding of several thousand cases. The first approach is excellent for careful doctrinal description. The second is necessary if the goal is to estimate temporal trends, model interaction effects, or train predictive models.

Choi describes this tradeoff bluntly: the most interesting empirical questions often require datasets that are too large for traditional hand-coding.[342] LLMs change the shape of that constraint. They do not eliminate error, but they make it possible to ask questions that simply could not be answered at scale fifteen years ago. The present study takes the second path. It accepts that some model-assisted codings will be wrong at the margin, and tries to manage that risk through prompts, validation, and human review, in exchange for a dataset large and detailed enough to make robust statistical analysis of DuPont practice possible.

[339] *Cf.* Choi, *supra* note 290, at 221–22 (discussing how published prompts and coding rules make LLM-based research more transparent than traditional hand-coding methods where inter-coder reliability is often assumed rather than demonstrated).
[340] Choi, *supra* note 290, at 216–19.
[341] Adhikary et al., *supra* note 314 (demonstrating that LLMs can extract structured legal attributes with high accuracy when evaluated against manually coded gold-standard annotations).
[342] Choi, *supra* note 290, at 214–15 ("[D]ramatic recent improvements in the performance of large language models (LLMs) now provide a potential alternative" to hand-coding, enabling empirical projects at scales previously impractical.).

C. Dataset Construction

1. Data Sources and Scope

This study analyzes published TTAB decisions from 2000 through 2025, drawn from the USPTO's TTAB Reading Room.[343] The study focused on inter partes proceedings (oppositions and cancellations) where likelihood of confusion under Section 2(d) constituted a primary issue.[344] Cases were identified using the Reading Room's metadata tags for proceeding type and legal grounds.

Inter partes proceedings involve adverse parties in an adversarial context resembling federal district court litigation, with pleadings, discovery, a trial period, and briefing.[345] This procedural posture makes them well-suited for studying how DuPont factors operate in contested cases. The dataset includes both precedential and non-precedential decisions to capture routine Board practice rather than only high-profile disputes.

Approximately 6,500 oppositions and 2,200 cancellation petitions are filed with the TTAB annually, though many settle or end in default without a merits decision.[346] The dataset captures only cases resulting in a published decision addressing likelihood of confusion.

2. Inclusion Criteria

To qualify for inclusion in the dataset, a decision must satisfy five criteria: (1) a published opinion, whether designated precedential or non-precedential;[347] (2) Section 2(d) likelihood of confusion as a substantive issue;[348] (3) a final merits determination rather than a procedural dismissal;[349] (4) a full DuPont factor analysis;[350] and (5) availability in machine-readable format.

The first criterion captures the Board's complete decisional output. Although non-precedential decisions lack binding authority, they constitute the vast majority of TTAB decisions and reflect actual adjudicatory practice.[351] Excluding them would introduce selection bias toward atypical cases that the Board deemed worthy of precedential designation.

The third and fourth criteria work in tandem: procedural dismissals for failure to prosecute, default, or lack of standing do not generate the substantive DuPont analysis that forms the basis of this study.[352] The final criterion reflects practical necessity; scanned images and

[343] *Decisions–TTAB Reading Room*, U.S. PAT. & TRADEMARK OFF., https://ttab-reading-room.uspto.gov/ (last visited Dec. 1, 2025).

[344] 15 U.S.C. § 1052(d).

[345] TRADEMARK TRIAL & APPEAL BD. MANUAL OF PROC. § 102.01 (June 2023) ("An inter partes proceeding before the Board is similar to a civil action in a federal district court.").

[346] *See Trademark Trial and Appeal Board Dashboard*, U.S. PAT. & TRADEMARK OFF., https://www.uspto.gov/dashboard/ttab/ (last visited Dec. 1, 2025).

[347] TBMP § 101.03 (June 2024) ("Since January 23, 2007, the Board has permitted citation to any Board decision or interlocutory order, although a decision or order designated as not precedential is not binding upon the Board, but may be cited for whatever persuasive value it might have.").

[348] 15 U.S.C. § 1052(d).

[349] *See* Hall & Wright, *supra* note 188, at 88–89 (emphasizing that inclusion criteria should identify decisions "that can answer the research question" and that "any selection criteria must be clearly articulated").

[350] *DuPont*, 476 F.2d at 1361 (establishing the thirteen-factor framework for likelihood of confusion analysis).

[351] *See* TBMP § 101.03 (June 2024); *see also In re* Soc'y of Health & Physical Educators, 127 U.S.P.Q.2d 1584, 1587 n.7 (T.T.A.B. 2018) ("Board decisions which are not designated as precedent are not binding on the Board, but may be cited and considered for whatever persuasive value they may have.").

[352] *Cf.* Beebe, *supra* note 29, at 1596–97 (describing inclusion criteria requiring "substantial use of a multi-factor test for the likelihood of consumer confusion" to ensure decisions contain the analytical content under study).

other non-machine-readable formats cannot be processed through the automated extraction pipeline.

3. Exclusion Criteria and Rationale

Three categories of cases were systematically excluded to ensure the dataset captured decisions where DuPont analysis actually determined outcomes.

a. Ex Parte Examination Appeals

Ex parte appeals from examining attorney refusals were excluded, despite constituting a substantial portion of the TTAB docket.[353] The analytical contexts differ too fundamentally to combine. Ex parte appeals pit applicant against examining attorney, with evidence limited to prosecution materials. Inter partes proceedings feature adverse parties, designated trial periods, and access to marketplace evidence that ex parte appellants can only dream about.[354]

The distinction runs deeper than procedure. Under the "Octocom rule," ex parte appeals analyze goods and services "as described in the application," while inter partes proceedings can consider commercial reality.[355] As former TTAB Judge Lorelei Ritchie explains, marketplace evidence "is less likely to be considered by the Board in [ex parte] likelihood of confusion cases, particularly with regard to the first through fourth du Pont factors."[356] Mixing proceeding types would introduce heterogeneity that statistical analysis cannot easily untangle.

b. Alternative Basis Decisions

Cases resolved on grounds other than likelihood of confusion were excluded. When the Board disposes of a proceeding on priority, standing, fraud, or another threshold ground, any confusion discussion becomes dicta.[357]

Priority illustrates the problem. An opposer must prove both priority and likelihood of confusion to prevail under Section 2(d).[358] When the Board finds no priority, it need not reach confusion at all, and any analysis it does provide reflects an alternative holding unconstrained by outcome-determinative rigor. The same logic applies to standing (now styled "entitlement to a statutory cause of action").[359] This exclusion ensures every coded decision reflects analysis that actually mattered.

c. Interlocutory Orders

Non-final decisions were excluded. Summary judgment denials identify disputed facts but resolve nothing; interlocutory rulings address procedure rather than substance.[360] Only final merits decisions contain the complete factor analysis this study requires.

4. Final Dataset Characteristics

After applying inclusion and exclusion criteria, the dataset comprised 3,999 TTAB inter partes decisions spanning 2000 through 2025. Some decisions involved multiple mark-to-mark comparisons, yielding approximately 4,500 total comparisons across 2,910 cases where

[353] *See* TBMP § 1201 (June 2024).

[354] *Compare* TBMP § 102.01 (June 2024) (inter partes proceedings "similar to a civil action in a federal district court"), *with* TBMP § 1203 (June 2024) (ex parte appeals "appellate in nature").

[355] *Octocom Sys., Inc. v. Houston Computs. Servs. Inc.*, 918 F.2d 937, 942 (Fed. Cir. 1990).

[356] LORELEI D. RITCHIE, *Recognizing the "Use"-fulness of Evidence at the TTAB*, 112 TRADEMARK REP. 635, 643 (2022).

[357] *See* TBMP § 309 (June 2024) (listing grounds for opposition and cancellation).

[358] *Empresa Cubana Del Tabaco v. Gen. Cigar Co.*, 753 F.3d 1270, 1275 (Fed. Cir. 2014).

[359] *Corcamore, LLC v. SFM, LLC*, 978 F.3d 1298, 1303 (Fed. Cir. 2020).

[360] *See* TBMP § 528 (June 2024); FED. R. CIV. P. 56(a).

likelihood of confusion was substantively decided.[361] Of these, roughly 2,400 resulted in findings of likely confusion while approximately 2,100 found no likelihood of confusion.

This near-even split was fortuitous but hardly surprising. As Priest and Klein demonstrated four decades ago, litigated cases cluster toward contested outcomes because parties with clearly losing positions settle rather than absorb the costs of proceeding.[362] Published TTAB decisions thus represent genuinely ambiguous disputes where both parties believed they had reasonable prospects of success. For studying how the Board applies DuPont factors, this selection effect proves advantageous: the dataset captures precisely those close cases where multifactor analysis theoretically matters most.

The temporal distribution proved relatively uniform, averaging approximately 160 decisions per year across the twenty-five-year period. This consistency enabled meaningful trend analysis to detect whether Board practice evolved over time. The sample size, at twelve times Beebe's 331 district court opinions, provided statistical power to identify effect sizes smaller than earlier trademark studies could reliably detect.[363]

The dataset's machine-readable format enabled LLM-based extraction at scale. TTAB opinions follow consistent structural conventions: procedural history, evidence discussion, factor-by-factor legal analysis, and conclusions.[364] This predictable format facilitated targeted data extraction focused on substantive likelihood-of-confusion determinations rather than procedural recitations or threshold issues. The result was a clean sample ideally suited for empirical analysis of *DuPont* factor application and outcome prediction.

What followed was less a validation than a reckoning. The thirteen factors went in; not all of them came out.

[361] Where an opposer asserted multiple prior registrations against an applied-for mark, each mark-to-mark comparison was analyzed separately. Multiple comparisons per decision do not create statistical dependency problems because the Board treats each comparison as analytically distinct.

[362]GEORGE L. PRIEST & BENJAMIN KLEIN, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1, 4–5 (1984) (demonstrating that rational litigants settle clear-cut cases, leaving genuinely contested disputes for adjudication and producing win rates that tend toward equilibrium).

[363]Beebe, *supra* note 29, at 1586–87 (analyzing 331 federal district court trademark opinions from 2000–2004). The present study's substantially larger sample enables detection of effects that would not reach statistical significance in smaller datasets.

[364]*See* Hall & Wright, *supra* note 188, at 67–69 (discussing the methodological advantages of coding judicial opinions with consistent structural formats).

PART III

THE DEATH OF DUPONT

I. Findings: The Two-Factor Reality

A. Thirteen Factors Collapse to Two

Analysis of nearly 4,000 TTAB decisions spanning twenty-five years yields an unambiguous conclusion: likelihood of confusion outcomes are overwhelmingly predictable from just two factors. Mark similarity (Factor 1) and goods/services relatedness (Factor 2) drive outcomes. The remaining eleven factors contribute virtually nothing to prediction accuracy.

1. The Core Finding

Using logistic regression, this study tested whether the thirteen-factor DuPont framework actually predicts outcomes or whether a simpler model performs comparably.[365] Model 1 used only Factors 1 and 2. Model 2 used all thirteen factors. The results were stark: the two-factor model achieved 99.37% classification accuracy; the thirteen-factor model achieved 99.79%.[366] Adding eleven factors improved accuracy by 0.42 percentage points. For every thousand cases decided, considering all thirteen factors instead of just two changes the predicted outcome in approximately four.

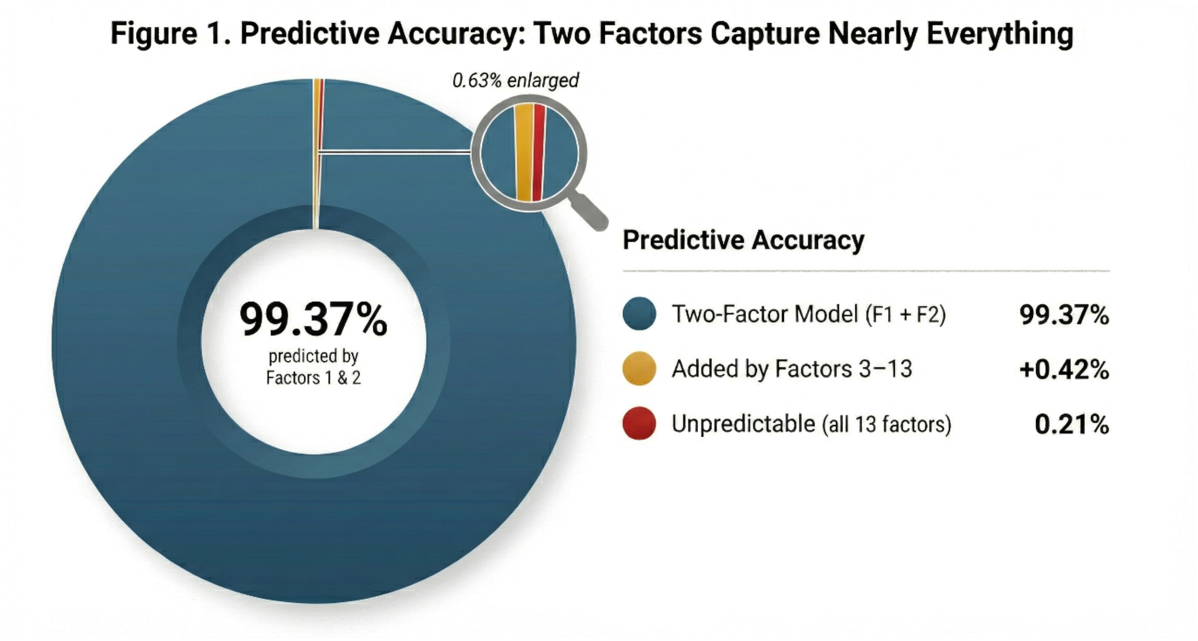


Figure 1. Predictive Accuracy: Two-Factor vs. Thirteen-Factor Model.
The two-factor model (Factors 1 and 2 only) achieves 99.37% accuracy in predicting TTAB likelihood of confusion outcomes. Adding the remaining eleven DuPont factors improves accuracy by only 0.42 percentage points.

The statistical measures reinforce this conclusion. The two-factor model's McFadden pseudo-R² is 0.947. McFadden himself characterized values between 0.2 and 0.4 as representing "excellent fit"; a value approaching 0.95 is virtually unprecedented in social science research.[367] The thirteen-factor model's pseudo-R² reaches 0.9925, an improvement of less

[365] Logistic regression is standard for binary outcome prediction in empirical legal studies. *See* LEE EPSTEIN & ANDREW D. MARTIN, AN INTRODUCTION TO EMPIRICAL LEGAL RESEARCH 234–42 (2014).

[366] Classification accuracy measures the percentage of cases in which the model correctly predicts the actual outcome. At 99.37%, the two-factor model misclassifies fewer than 1 in 150 cases.

[367] Daniel McFadden, *Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments*, *in* BEHAVIOURAL TRAVEL MODELLING 279, 306–07 (David A. Hensher & Peter R.

than five percent. The Area Under the Curve (AUC-ROC), measuring discrimination ability, reaches 0.9984 for the two-factor model and 1.0 for the full model.[368] By every metric designed to assess predictive power, Factors 1 and 2 capture nearly all the information contained in the complete DuPont framework.

2. Situating the Finding in Prior Research

This finding aligns with prior empirical work while dramatically sharpening its conclusions. In 2006, Barton Beebe analyzed 331 federal district court trademark opinions and found that mark similarity and goods relatedness dominated outcomes while peripheral factors "stampeded" to conform to the ultimate conclusion.[369] The present study, with a sample twelve times larger and using scaled intensity coding rather than binary variables, confirms Beebe's intuition with statistical precision.

More recently, Daryl Lim's analysis of appellate trademark decisions identified a "potent trio" of factors guiding judicial outcomes: actual confusion, mark similarity, and competitive proximity.[370] Lim observed that courts "economize" by analyzing only a handful of factors and "fold" related factors into one another.[371] The regression analysis confirms mark similarity and competitive proximity (goods relatedness) as the predictive core. Actual confusion, Lim's third member of the trio, proves determinative when present but rarely appears at the registration stage. In the TTAB dataset, only 16.3% of all comparisons had actual confusion evidence that weighted in either direction. The true core of TTAB predictability consists of Factors 1 and 2 alone.

3. Factor-by-Factor Analysis

Figure 2 presents these results as a volcano plot, a visualization technique borrowed from genomics that displays each factor's effect size (multivariate regression coefficient) against its statistical significance ($-\log_{10}$ of the p-value).[372] Bubble size represents how frequently each factor appears in the dataset, ranging from 153 observations (Factor 11) to 2,875 observations (Factor 1). The visual pattern is striking.

Factors 1 and 2 occupy isolated positions in the upper-right quadrant. Factor 1 (mark similarity) produces a coefficient of 1.82 ($p < 10^{-19}$); Factor 2 (goods relatedness) produces a coefficient of 1.62 ($p < 10^{-14}$). These are large effects with overwhelming statistical significance. Notably, these two factors are also among the most frequently analyzed, each appearing in over 2,800 decisions. The factors that matter most are the factors the Board examines most often.

Stopher eds., 1979) ("[V]alues of 0.2 to 0.4 for ρ^2 represent excellent fit."). McFadden's pseudo-R^2 is the standard goodness-of-fit measure for logistic regression models.

[368] AUC-ROC (Area Under the Receiver Operating Characteristic Curve) measures a model's ability to discriminate between positive and negative cases across all possible classification thresholds. A value of 1.0 represents perfect discrimination. *See* Tom Fawcett, *An Introduction to ROC Analysis*, 27 PATTERN RECOGNITION LETTERS 861 (2006).

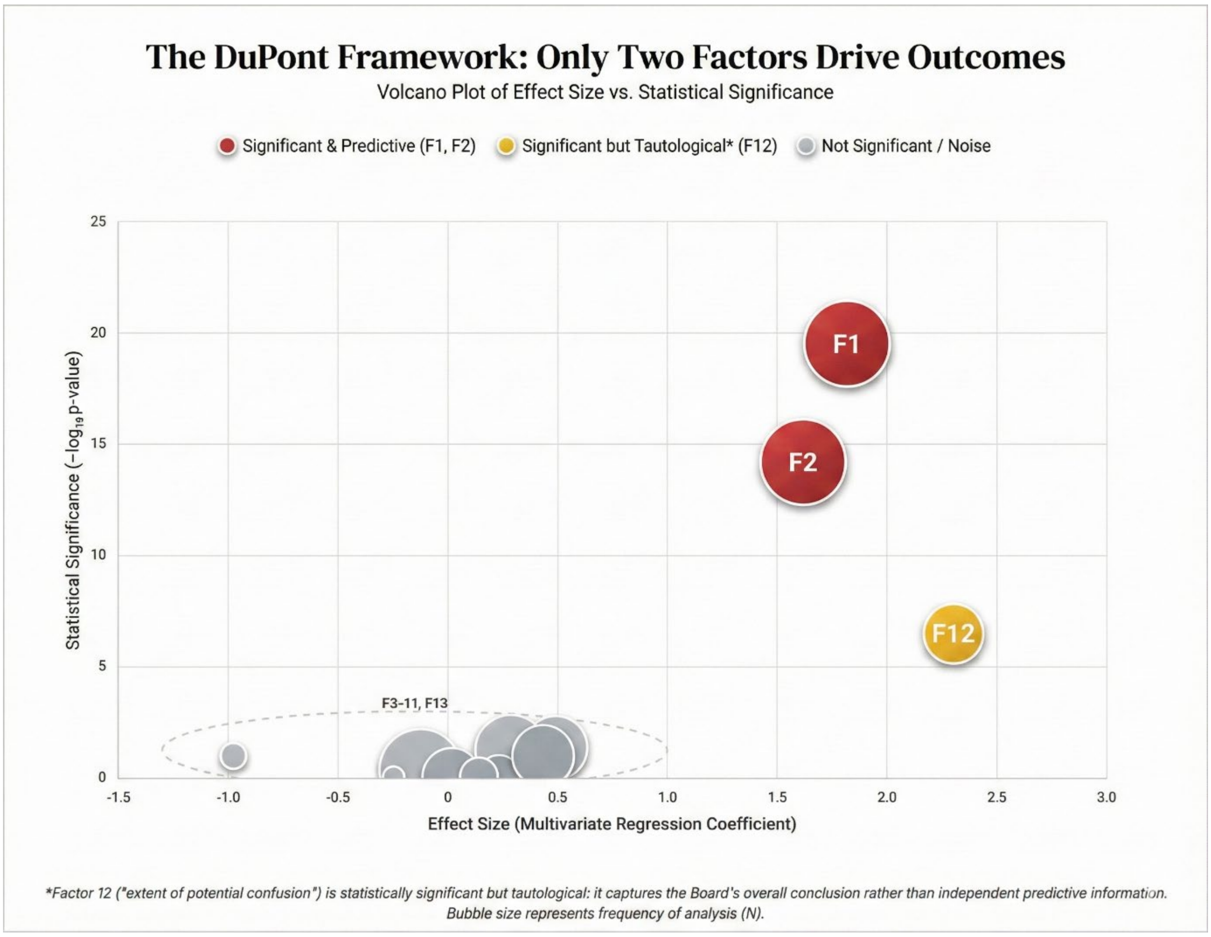
[369] Beebe, *supra* note 29, at 1628–31 (describing "stampeding" as the phenomenon whereby factors beyond the core predictors align with the ultimate outcome rather than independently influencing it).

[370] Lim, *Trademark Confusion Revealed*, *supra* note 190, at 1290.

[371] *Id*. at 1291-92.

[372] Volcano plots simultaneously display effect magnitude and statistical significance, enabling rapid identification of variables that exhibit both large effects and high confidence. The technique is standard in differential expression analysis. *See* Wei Li, *Volcano Plots in Analyzing Differential Expressions with mRNA Microarrays*, 10 J. BIOINFORMATICS & COMPUTATIONAL BIOLOGY 1231001 (2012).

The remaining factors tell a different story. They cluster along the bottom of the plot, hugging the x-axis in a mass of statistical insignificance. Factor 4 (purchaser sophistication) shows a coefficient of 0.34 ($p = 0.13$). Factor 7 (actual confusion) shows a coefficient of 0.02 ($p = 0.96$). Factor 8 (concurrent use) shows a coefficient of 0.23 ($p = 0.77$). Factor 11 (right to exclude) shows a coefficient of −0.25 ($p = 0.87$).[373] These factors contribute nothing systematic to outcome prediction. Their coefficients are statistical noise.



The DuPont Framework: Only Two Factors Drive Outcomes

Volcano Plot of Effect Size vs. Statistical Significance

*Factor 12 ("extent of potential confusion") is statistically significant but tautological: it captures the Board's overall conclusion rather than independent predictive information. Bubble size represents frequency of analysis (N).

Figure 2. The DuPont Framework: Only Two Factors Drive Outcomes.

Volcano plot displaying effect size (multivariate regression coefficient) on the x-axis against statistical significance (−log₁₀ p-value) on the y-axis for all thirteen DuPont factors. Bubble size represents frequency of analysis (n). Red bubbles indicate factors that are both statistically significant and substantively predictive (Factors 1 and 2). The gold bubble indicates Factor 12, which achieves statistical significance but functions tautologically. Gray bubbles indicate factors that fail to reach statistical significance. Factors 1 and 2 are isolated in the upper-right quadrant; the remaining factors cluster along the bottom of the plot.

This pattern holds regardless of sample size. Factor 3 (trade channels) appears in 2,659 decisions, making it one of the most frequently analyzed factors in the dataset. Yet it produces a coefficient of −0.12 ($p = 0.46$) when Factors 1 and 2 are controlled. The same is true for Factor 5 (mark fame, $n = 2,058$, $p = 0.06$) and Factor 13 (other probative facts, $n = 1,478$, $p = 0.10$). Exposed to a multivariate test, these factors reveal themselves as what Beebe suspected and Lim confirmed: redundant proxies that courts "fold" into mark

[373] None of these p-values approaches conventional significance thresholds ($p < 0.05$ or even $p < 0.10$).

similarity and goods relatedness rather than independent predictors of confusion.[374] The separate factors are not independent measurements; they are proxies for the same underlying constructs. The appearance of comprehensive analysis masks redundancy.

Only Factor 12 achieves statistical significance beyond Factors 1 and 2, shown in gold on the plot (coefficient 2.30, $p < 0.001$). But this result is tautological rather than informative. Factor 12 instructs decisionmakers to assess "the extent of potential confusion, i.e., whether de minimis or substantial."[375] It essentially requests a bottom-line judgment that incorporates all other considerations.[376] Factor 12 correlates with outcomes not because it measures something independent but because it encapsulates the Board's holistic conclusion. It is the outcome wearing a factor's mask.

4. Implications

The practical implication is stark. Likelihood of confusion at the TTAB is predictable from mark similarity and goods relatedness. If the marks are similar and the goods overlap, confusion will be found. If either element is absent, confusion will not be found. Edge cases exist, but they represent fewer than one percent of outcomes.

Some readers may object that courts must have reasons for discussing all thirteen factors. This objection conflates rhetoric with reality. Courts discuss Factor 8 (concurrent use) because doctrine requires it, not because it changes outcomes.[377] Parties brief Factor 5 (fame) because the framework invites it, not because fame independently predicts results once mark similarity is controlled. When the analysis tests which factors actually predict which party wins, the answer is unambiguous: Factors 1 and 2 predict; the others do not.

The DuPont framework, celebrated for fifty years as a flexible, comprehensive approach to likelihood of confusion analysis, collapses empirically to a two-factor test.[378] The remaining eleven factors are an expensive ritual serving no systematic predictive function. The emperor, it turns out, has been wearing a considerably simpler outfit all along.

B. The Goldilocks Zone: Visualizing Two-Factor Dominance

The regression analysis establishes statistical dominance; a visualization makes it unmistakable. Figure 3 presents a heatmap plotting confusion outcomes by Factor 1 (mark similarity) and Factor 2 (goods relatedness) scores across 2,835 TTAB decisions where the

[374] Beebe found that judges "stampede" non-dispositive factors "to conform to the test outcome." Beebe, *supra* note 29, at 1582. Lim confirmed that courts "economize" by analyzing only a handful of factors and "fold" others into those core considerations. Lim, *Trademark Confusion Revealed*, *supra* note 190, at 1291–92
The present multivariate results provide statistical confirmation: factors that appear significant in isolation lose significance when mark similarity and goods relatedness are controlled, indicating that their apparent predictive power derives from correlation with Factors 1 and 2 rather than independent contribution.

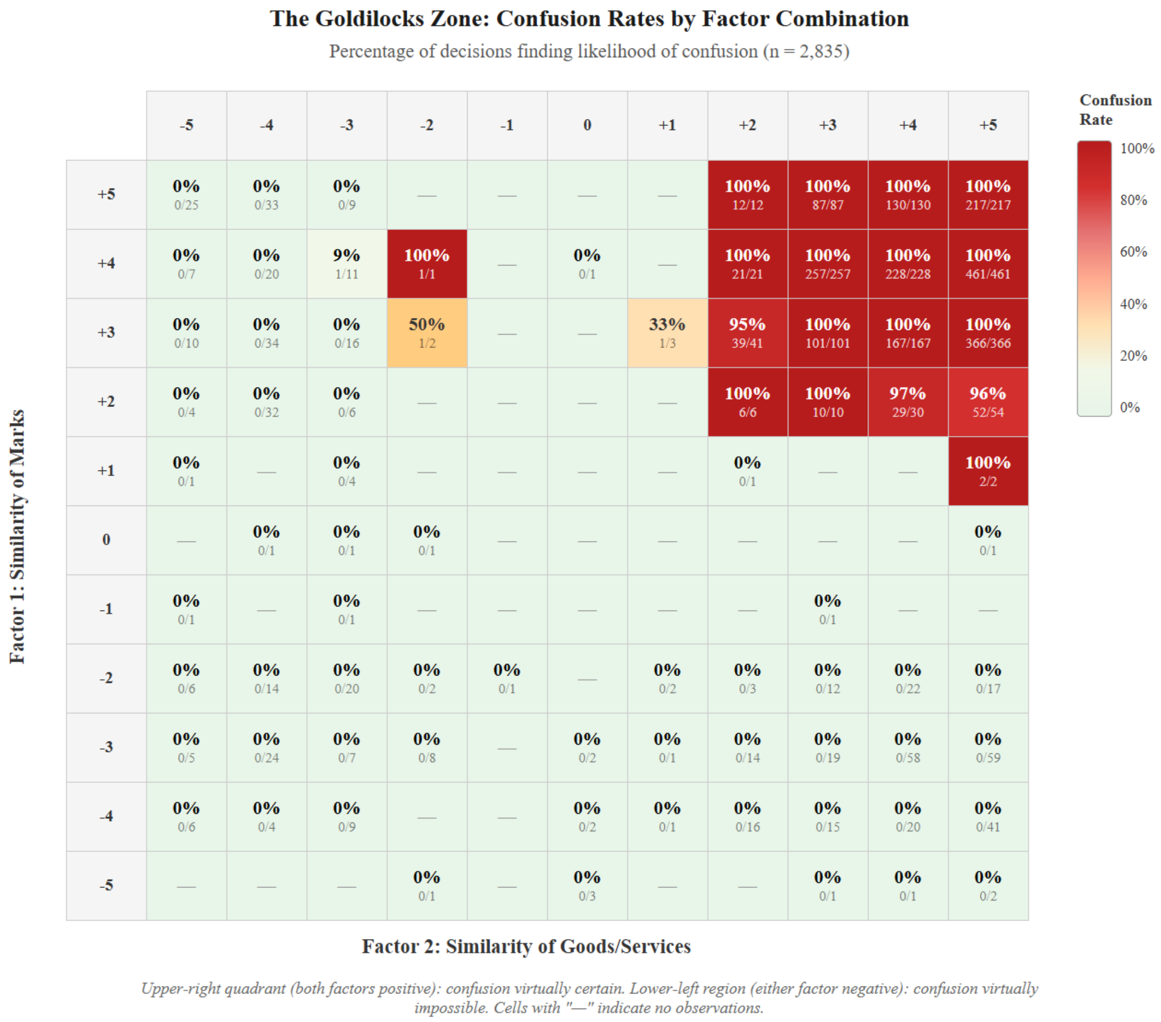
[375] *DuPont*, 476 F.2d at 1361.

[376] Factor 13 similarly instructs consideration of "[a]ny other established fact probative of the effect of use." *Id.* Both factors function as catch-alls that absorb holistic assessment rather than measure discrete phenomena.

[377] The Board routinely recites that it has "considered all *DuPont* factors for which there is evidence and argument" before focusing its analysis on the factors that actually matter. *See, e.g.*, *Guild Mortg.*, 912 F.3d 1376, 1379 (Fed. Cir. 2019).

[378] *DuPont*, 476 F.2d at 1361 (articulating the thirteen-factor framework).

Board conducted substantive DuPont analysis of both factors and reached a merits determination on likelihood of confusion.[379]



The Goldilocks Zone: Confusion Rates by Factor Combination

Percentage of decisions finding likelihood of confusion (n = 2,835)

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
+5	0% 0/25	0% 0/33	0% 0/9	—	—	—	—	100% 12/12	100% 87/87	100% 130/130	100% 217/217
+4	0% 0/7	0% 0/20	9% 1/11	100% 1/1	—	0% 0/1	—	100% 21/21	100% 257/257	100% 228/228	100% 461/461
+3	0% 0/10	0% 0/34	0% 0/16	50% 1/2	—	—	33% 1/3	95% 39/41	100% 101/101	100% 167/167	100% 366/366
+2	0% 0/4	0% 0/32	0% 0/6	—	—	—	—	100% 6/6	100% 10/10	97% 29/30	96% 52/54
+1	0% 0/1	—	0% 0/4	—	—	—	—	0% 0/1	—	—	100% 2/2
0	—	0% 0/1	0% 0/1	0% 0/1	—	—	—	—	—	—	0% 0/1
-1	0% 0/1	—	0% 0/1	—	—	—	—	—	0% 0/1	—	—
-2	0% 0/6	0% 0/14	0% 0/20	0% 0/2	0% 0/1	—	0% 0/2	0% 0/3	0% 0/12	0% 0/22	0% 0/17
-3	0% 0/5	0% 0/24	0% 0/7	0% 0/8	—	0% 0/2	0% 0/1	0% 0/14	0% 0/19	0% 0/58	0% 0/59
-4	0% 0/6	0% 0/4	0% 0/9	—	—	0% 0/2	0% 0/1	0% 0/16	0% 0/15	0% 0/20	0% 0/41
-5	—	—	—	0% 0/1	—	0% 0/3	—	—	0% 0/1	0% 0/1	0% 0/2

Factor 1: Similarity of Marks
Factor 2: Similarity of Goods/Services
Confusion Rate
100%
80%
60%
40%
20%
0%

Upper-right quadrant (both factors positive): confusion virtually certain. Lower-left region (either factor negative): confusion virtually impossible. Cells with "—" indicate no observations.

Figure 3. The Goldilocks Zone: Confusion Rates by Factor Combination.
Cell color represents the percentage of decisions finding likelihood of confusion for each Factor 1/Factor 2 score combination. Dark red indicates confusion found in nearly all cases; light green indicates confusion rarely or never found. Each cell displays both the confusion rate and the underlying observation count (n/N). Gray cells indicate no observations for that combination. N = 2,835.

1. A Binary World

The image reveals a stark binary pattern. There is no gradient. There is no middle ground. There is only the Goldilocks Zone and everywhere else (with outliers!).

The upper-right quadrant glows red. When marks are similar and goods overlap, confusion is virtually certain. At maximum scores (Factor 1 = +5, Factor 2 = +5), all 217 cases found confusion. At Factor 1 = +4 and Factor 2 = +5, all 461 cases found confusion. Across the quadrant where both factors score +2 or higher, confusion rates exceed 95%.[380] This is the

[379] This subset excludes cases where the Board did not reach a merits determination on likelihood of confusion (e.g., procedural dismissals, consent agreements), cases resolved on alternative grounds, and cases where either Factor 1 or Factor 2 was not substantively analyzed.

[380] The only exceptions within this quadrant involve the weak mark and crowded field cases discussed below. *See infra* notes 382–84 and accompanying text.

Goldilocks Zone: conditions "just right" for a confusion finding. The Board almost never says no.

Everywhere else is a sea of green. When either factor scores negative, confusion rates plummet to zero. Dissimilar marks with identical goods (Factor 1 = -3, Factor 2 = +5): zero of 59 cases found confusion. Similar marks with unrelated goods (Factor 1 = +5, Factor 2 = -4): zero of 33 cases found confusion.[381] The two factors operate conjunctively, like a two-key system for launching missiles. Both keys must turn. A negative score on either vetoes confusion regardless of the other's value. Of 847 decisions where Factor 1 scored negative, exactly zero found likelihood of confusion. Not one. The pattern admits no exceptions.

What happened to the other eleven factors? They appear in the decisions. They fill pages of analysis. But when both keys have turned, they do not stop the launch. And when either key remains unturned, they rarely start it. Factors 3 through 13 are the procedural equivalent of decorative columns: they look structural but bear no weight.

2. Outliers: Cases for Future Study

A handful of outliers exist on both sides of the boundary. Five decisions found no confusion despite falling within the Goldilocks Zone. Each involved either a conceptually weak mark or a crowded field. In *Box, Inc. v. Ikbariyeh*, the Board found BOX conceptually weak for cloud storage services given widespread third-party use in the industry.[382] In *El Burro, Inc. v. Knuckle Sandwich LLC*, eleven third-party uses of "El Burro" for Mexican restaurants convinced the Board that the mark lacked distinctiveness, finding Factor 6 "dispositive" despite identical services.[383] In *IAC Search & Media, Inc. v. ASKBOT*, the Board found ASK weak and descriptive for question-and-answer software.[384]

Six decisions found confusion despite falling outside the Goldilocks Zone. Each involved exceptional fame bridging a goods gap or a junior mark incorporating a senior mark entirely. In *Hasbro, Inc. v. Braintrust Games, Inc.*, the "well-known and strong" CLUE mark for board games supported a confusion finding against NO FRIGGIN CLUE despite only moderate mark similarity (Factor 1 = +1).[385] In *Trek Bicycle Corp. v. Natural Balance Foods Ltd.*, the Board found confusion likely between TREK for bicycles and TREK for snack bars (Factor 2 = -3), reasoning that consumers encountering the snack bars in bike shops would assume sponsorship.[386] And in *Recot, Inc. v. Becton*, the Federal Circuit vacated the Board's finding

[381] This pattern holds across the entire dataset. Of 847 decisions where Factor 1 scored negative, exactly zero found confusion regardless of Factor 2's value.

[382] *Box, Inc. v. Hakem Ikbariyeh*, Cancellation No. 91202576, 2016 WL 3647918 (T.T.A.B. July 7, 2016) (non-precedential).

[383] *El Burro, Inc. v. Knuckle Sandwich LLC*, Cancellation No. 92075933, 2023 WL 3662417 (T.T.A.B. May 26, 2023) (non-precedential) (finding that the crowded field "outweighs the other factors that favor likelihood of confusion").

[384] *IAC Search & Media, Inc. v. ASKBOT, Spa*, Cancellation No. 92060041, 2018 WL 4215648 (T.T.A.B. Aug. 31, 2018) (non-precedential).

[385] *Hasbro, Inc. v. Braintrust Games, Inc.*, Opposition No. 91169603, 2009 WL 2595248 (T.T.A.B. Aug. 24, 2009) (non-precedential) (finding marks only "somewhat, but not strongly, similar" yet sustaining opposition based on CLUE's renown for identical goods).

[386] *Trek Bicycle Corp. v. Natural Balance Foods Ltd.*, Opposition No. 91221706, 2019 WL 1172919 (T.T.A.B. Mar. 13, 2019) (non-precedential).

of no confusion between FRITO-LAY and FIDO LAY for pet treats, holding that the Board had improperly discounted the FRITO-LAY mark's exceptional fame.[387]

These outlier cases merit individual study. They represent the rare circumstances where Factors 5 (fame) or 6 (crowded field) genuinely moved the needle, and future research should examine whether they reflect principled exceptions or simply noise. But the heatmap's overwhelming message is conformity to a two-dimensional pattern. The Board's actual decision rule reduces to two questions: Are the marks similar? Are the goods related? If yes to both, find confusion. If no to either, don't. The remaining eleven factors provide rhetorical scaffolding for conclusions the first two factors have already determined.

C. The Categorical Collapse: A Decision Rule That Outperforms Statistics

The heatmap's stark binary pattern suggests something beyond mere statistical correlation. It suggests a categorical decision rule. To test this hypothesis, I applied the simplest possible classification model: a 2x2 decision matrix that predicts confusion if and only if both Factor 1 and Factor 2 favor confusion.

	F2 Favors Confusion	***F2 Disfavors Confusion***
F1 Favors Confusion	Predict: CONFUSION	Predict: No Confusion
F1 Disfavors Confusion	Predict: No Confusion	Predict: No Confusion

This rule achieves 99.52% accuracy across 4,757 individual trademark comparisons. Only 23 cases deviate from the pattern. More striking still, this categorical rule *outperforms* logistic regression models. The breakdown by category reveals why:

Category	***Total N***	***Confusion***	***No Confusion***	***Predicted***	***Correct***	***Accuracy***
Both Favor	3,405	3,386	19	Confusion	3,386	99.44%
Both Disfavor	277	0	277	No Confusion	277	100.00%
F1 Favor, F2 Disfavor	537	3	534	No Confusion	534	99.44%
F1 Disfavor, F2 Favor	538	1	537	No Confusion	537	99.81%

When both factors favor confusion (72% of comparisons), the Board finds confusion in 99.44% of cases. When both factors disfavor confusion (6% of comparisons), the Board finds no confusion in 100% of cases. And when the factors point in opposite directions (22% of comparisons), the Board finds no confusion in all but four cases.

That a categorical rule outperforms continuous regression models carries theoretical implications. The Board's reasoning appears to be threshold-based, not proportional. A mark is either "similar enough" or not. Goods are either "related enough" or not. The degree of similarity beyond a certain threshold doesn't proportionally increase confusion likelihood. This explains the heatmap's abrupt transitions: there is no gradient because the underlying decision process admits none.

The rule's four false negatives merit attention. Three involved Factor 1 favoring confusion but Factor 2 disfavoring it; one involved the reverse. Each case involved exceptional

[387] *Recot, Inc. v. M.C. Becton*, 214 F.3d 1322, 1327–28 (Fed. Cir. 2000) (vacating and remanding where Board failed to accord proper weight to fame of FRITO-LAY mark in analyzing likelihood of confusion with FIDO LAY for pet treats).

circumstances: famous marks bridging goods gaps or conceptual distinctions overriding surface similarity. These comparisons are associated with the outlier cases identified in the previous section. The rule's 19 false positives all fell within the Goldilocks Zone but involved weak marks or crowded fields. Again, these are the same outliers.

What emerges is a picture of the TTAB's confusion analysis as an almost mechanical process. The Board applies a binary test disguised as a multifactor balancing framework. When counsel brief all thirteen DuPont factors and judges discuss each at length, they are performing an elaborate ritual whose outcome was determined the moment the Board assessed whether the marks were similar and the goods related. The other eleven factors are not weights in a balance, but commentary on a conclusion already reached.

D. Why Courts Maintain the Fiction

The empirical evidence is overwhelming: likelihood of confusion outcomes depend almost entirely on two factors, not thirteen. Yet courts continue to invoke the full DuPont framework in virtually every case, ritualistically discussing factors that contribute nothing to the result. Why does the fiction persist?

The answer lies in a combination of cognitive limitations and institutional incentives. Courts aren't intentionally deceiving litigants - they're not huddled in chambers, cackling over their elaborate charade. They're operating within a system that makes abandoning the comprehensive framework difficult, even when the data proves it dysfunctional.

1. Why Judges Believe It Works

Judges genuinely believe they're weighing all relevant factors. This belief reflects a well-documented cognitive phenomenon: post-hoc rationalization.[388] Decision-makers reach conclusions based on limited information, then construct elaborate justifications that appear to consider many variables.[389] The reasoning comes after the decision, not before it. In trademark cases, judges likely form preliminary views based on mark similarity and goods relatedness (the two factors most salient and easiest to assess) then write opinions discussing all thirteen factors to justify conclusions already reached.[390]

Confirmation bias reinforces this pattern.[391] Once a judge determines that marks are similar and goods overlap, peripheral factors are interpreted to support that assessment. Beebe's 2006 study called this "stampeding": judges march the remaining factors into line with the outcome the core factors dictate.[392] A finding that purchasers are "ordinarily sophisticated" might cut against confusion when the judge has already concluded confusion unlikely, but the same finding gets dismissed as insufficient protection when marks and goods point the other

[388] DAN SIMON, *A Third View of the Black Box: Cognitive Coherence in Legal Decision Making*, 71 U. CHI. L. REV. 511, 520–25 (2004) (describing "coherence-based reasoning" in which decision-makers unconsciously adjust their assessments of ambiguous evidence to support emerging conclusions).

[389] Richard E. Nisbett & Timothy D. Wilson, *Telling More Than We Can Know: Verbal Reports on Mental Processes*, 84 PSYCHOL. REV. 231, 233–35 (1977) (demonstrating that individuals lack reliable introspective access to their actual decision-making processes and instead construct plausible narratives *post hoc*).

[390] *See* Jonathan Haidt, *The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment*, 108 PSYCHOL. REV. 814, 818–20 (2001) (arguing that moral reasoning typically serves to justify intuitive judgments rather than to reach them); Simon, *supra* note 388, at 537–38.

[391] Raymond S. Nickerson, *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*, 2 REV. GEN. PSYCHOL. 175, 177–80 (1998) (surveying extensive psychological literature on the tendency to seek, interpret, and recall information in ways that confirm pre-existing beliefs).

[392] Beebe, *supra* note 29, at 1649–51 (coining the term "stampeding" to describe the phenomenon in which once a court determines outcomes on core factors, "the remaining factors tend to fall into line behind them").

way.[393] The DuPont factors become tools for rationalization rather than genuine inputs affecting decisions. They are, in effect, judicial window dressing, albeit very thorough window dressing.

This isn't a character flaw. It's how human cognition works. Judges, like everyone else, have limited capacity to process multiple variables simultaneously.[394] When faced with complex multifactor tests, decision-makers rely on heuristics, mental shortcuts that prioritize the most obviously relevant considerations.[395] Kahneman calls this System 1 thinking: fast, intuitive, and dominant.[396] In trademark law, mark similarity and goods relatedness are concrete and legally central. Other factors require inferential reasoning from limited evidence. Judges naturally gravitate toward what matters most, but the comprehensive framework creates an illusion that all factors receive equal consideration.[397]

The structure of TTAB opinions reinforces this illusion. Decisions follow a predictable template: recite all thirteen DuPont factors, discuss each factor for which evidence exists, conclude.[398] This ritualistic format signals thoroughness and compliance with precedent.[399] But format doesn't equal substance. A forty-page opinion discussing all thirteen factors reaches the same conclusion as a five-page opinion focusing only on Factors 1 and 2 because the outcome was determined by those two factors alone. The eleven additional factors are, to borrow a phrase, sound and fury signifying nothing.

2. Why the Framework Persists Anyway

Even if judges suspected the truth, institutional pressures would keep them performing the ritual. Appellate review focuses on whether the lower court "considered all relevant factors."[400] An opinion omitting discussion of a DuPont factor, even one with no evidentiary support, risks reversal for "failing to apply the correct legal standard."[401] In *In re Guild Mortgage*, the Federal Circuit vacated a TTAB decision for the sin of not addressing factor eight, even though the Board's opinion indicated it had "considered the factors for which

[393] *Compare Majestic Distilling*, 315 F.3d at 1317 (finding sophisticated purchasers insufficient to avoid confusion where marks and goods were similar), *with Shell Oil*, 992 F.2d at 1208 (emphasizing purchaser sophistication as weighing against confusion where goods differed).

[394] Chris Guthrie, Jeffrey J. Rachlinski & Andrew J. Wistrich, *Inside the Judicial Mind*, 86 CORNELL L. REV. 777, 784–88 (2001) (presenting empirical study of 167 federal magistrate judges demonstrating susceptibility to cognitive illusions including anchoring, framing effects, and hindsight bias).

[395] Gerd Gigerenzer & Wolfgang Gaissmaier, *Heuristic Decision Making*, 62 ANN. REV. PSYCHOL. 451, 454–58 (2011) (explaining how heuristics enable efficient decision-making by focusing on the most diagnostic cues while ignoring less relevant information).

[396] DANIEL KAHNEMAN, THINKING, FAST AND SLOW 19–30 (2011) (distinguishing between intuitive "System 1" processing, which operates automatically and quickly, and deliberative "System 2" processing, which requires conscious effort).

[397] *See* Beebe, *supra* note 29, at 1581–82 (noting that the multifactor test creates "an impression of rigor and comprehensiveness" that may not reflect actual decision-making processes).

[398] *See, e.g.*, *DuPont*, 476 F.2d at 1361 (establishing that "each" of the thirteen factors "must be considered" when relevant evidence exists); *Dixie*, 105 F.3d at 1406–07 (reaffirming that the *DuPont* factors "must be considered" in every case, though noting that "not all of the *DuPont* factors are relevant or of similar weight in every case").

[399] *See* Frederick Schauer, *Formalism*, 97 YALE L.J. 509, 510–15 (1988) (analyzing how formal legal structures constrain and channel judicial reasoning, creating predictability even when the formal categories do not map onto substantive considerations).

[400] *See DuPont*, 476 F.2d at 1361 ("[I]n every case turning on likelihood of confusion, it is the duty of the examiner, the board and this court to find, upon consideration of all the evidence, whether or not confusion appears likely.").

[401] *See Dixie*, 105 F.3d at 1406 ("In discharging this duty, the thirteen *DuPont* factors 'must be considered' 'when [they] are of record.'" (quoting *DuPont*, 476 F.2d at 1361)).

there was argument and evidence."[402] The Board's actual error was failing to *discuss* the factor, which is a distinction that makes sense only if one believes the ritual matters independent of the result.

This creates a perverse incentive structure. Discussing all thirteen factors is safer than honestly acknowledging that only two matter.[403] The cost of comprehensiveness is judicial time. However that cost is diffuse, spread invisibly across thousands of opinions. The cost of appearing to take shortcuts is reversal in the specific case, which reflects poorly on the individual judge.[404] Rational judges, facing this asymmetric payoff, opt for comprehensive ritual over honest parsimony.[405] One might call this the "cover your factors" strategy.

Strategic considerations compound the problem. Doctrinal ambiguity serves judicial interests that clarity would not.[406] Flexibility in factor weighting allows judges to reach desired outcomes without committing to rules that would bind future cases.[407] If courts explicitly acknowledged that Factors 1 and 2 determine 95% of outcomes, they would face pressure to explain the exceptional 5% and to develop clear doctrine for when peripheral factors actually matter. The current framework avoids this obligation by treating every case as sui generis, a unique snowflake of trademark confusion that defies systematic analysis.[408]

Finally, there is simple inertia. DuPont is over fifty years old.[409] Thousands of decisions cite it.[410] The Federal Circuit is bound by its own precedent, and absent en banc reconsideration or Supreme Court intervention, lower courts cannot abandon the framework even if persuaded by empirical evidence.[411] Legal doctrine exhibits powerful path dependence: once

[402] *Guild Mortg.*, 912 F.3d at 1380 ("The Board's opinion . . . does not mention factor 8, let alone address Guild's argument and evidence directed to that factor. The Board erred in failing to consider Guild's arguments and evidence."). The Guild Mortgage applicant and registrant had coexisted for over forty years without any evidence of actual confusion, a fact the Federal Circuit deemed sufficiently important that its omission from discussion warranted vacatur and remand. *Id.*

[403] *Cf.* RICHARD A. POSNER, HOW JUDGES THINK 61 (2008) (observing that judges have "a healthy aversion to appellate reversal").

[404] *See* LEE EPSTEIN, WILLIAM M. LANDES & RICHARD A. POSNER, THE BEHAVIOR OF FEDERAL JUDGES 52–53 (2013) (examining judges' aversion to reversal as a key variable in judicial behavior and finding that trial judges adjust behavior to minimize reversal risk).

[405] *See id.* at 46 (arguing that judges are rational actors who respond to incentive structures in predictable ways). The authors characterize judges as "labor-market participants" whose behavior is "shaped by the conditions and incentives of their employment." *Id.* at 2.

[406] *Cf.* Schlag, *supra* note 164, at 400–06 (observing that flexible standards confer discretion that serves institutional interests unavailable under rigid rules).

[407] *See Shell Oil*, 992 F.2d at 1206 (noting that *DuPont* factors "may play more or less weighty roles in any particular determination"); *In re* Nat'l Data Corp., 753 F.2d 1056, 1058 (Fed. Cir. 1985) (affirming that decision-makers may give "more or less weight" to particular features of trademarks). This flexibility is presented as doctrinal virtue rather than analytical vice.

[408] *Cf.* Frederick Schauer, *Precedent*, 39 STAN. L. REV. 571, 595–97 (1987) (discussing how treating cases as unique undermines the constraint function of precedent while preserving its legitimating appearance).

[409] *DuPont* was decided October 17, 1973. *DuPont*, 476 F.2d 1357.

[410] *See* McCarthy, *supra* note 33, § 24:30 (describing *DuPont* as the "leading case" applied in "thousands" of TTAB and Federal Circuit decisions); *see also supra* Part I.B (describing dataset of over 10,000 TTAB decisions applying *DuPont*).

[411] *See South Corp.*, 690 F.2d at 1370 (holding that Federal Circuit panels are bound by prior panel decisions absent *en banc* reconsideration or intervening Supreme Court authority).

established, rules persist even when their original justifications have long since eroded.[412] The common law, as Hathaway observes, is "firmly guided by the heavy hand of the past."[413]

The legal profession has adapted accordingly. Practitioners know how to litigate thirteen-factor cases. Forms, practice guides, and CLE materials teach comprehensive DuPont analysis.[414] Changing the framework would require re-educating the trademark bar, revising practice materials, and adjusting litigation strategies that have been refined over decades.[415] This collective investment in the status quo creates resistance to reform, even reform that would benefit everyone by eliminating eleven steps of pointless analysis.

Beebe's 2006 study provided suggestive evidence of the framework's dysfunction, but its sample size and binary coding limited its persuasiveness.[416] The present study offers definitive proof at a scale that should end the debate. But proof is not self-executing. The lag between empirical discovery and doctrinal reform can span decades, as the legal academy learns at one speed while courts move at quite another.[417] The death of DuPont's thirteen-factor comprehensiveness is now empirically established. Its doctrinal burial awaits a Federal Circuit willing to stop pretending.

E. The Real Cost of Pretending

If the DuPont framework were merely academic inefficiency, perhaps it could be tolerated. But the gap between what trademark law claims to do and what it actually does imposes real costs on real parties, both economic and institutional.

1. Economic Waste and Distributional Harm

The framework's primary economic cost is straightforward: it requires parties to litigate factors that do not matter. If two factors predict outcomes, resources expended on the other eleven represent pure waste. Parties commission expert reports on purchaser sophistication that adjudicators ignore; they develop evidence of intent that proves immaterial; they brief factors whose resolution will not affect the result. One might call this the "thirteen-factor tax," payable regardless of relevance.

The tax is not trivial. Median trademark litigation costs range from $150,000 to $400,000 depending on the stakes, with contested cases routinely exceeding $600,000 through trial.[418] Confusion surveys alone run $30,000 to $150,000.[419] Parties thus spend substantial sums

[412] Hathaway, *supra* note 149, at 604 ("The doctrine of *stare decisis* . . . creates an explicitly path-dependent process. Later decisions rely on, and are constrained by, earlier decisions.").

[413] *Id.* at 643.

[414] *See, e.g.*, TRADEMARK MANUAL OF EXAMINING PROCEDURE § 1207.01 (Oct. 2023) (organizing likelihood of confusion analysis around *DuPont* factors); RICHARD L. KIRKPATRICK, LIKELIHOOD OF CONFUSION IN TRADEMARK LAW §§ 4:1–4:13 (2024) (devoting extensive treatment to each factor).

[415] *See* Hathaway, *supra* note 149, at 628–30 (observing that path dependence generates "switching costs" that create resistance to doctrinal change even when change would be efficient).

[416] Beebe, *supra* note 29, at 1591–96 (studying 331 federal district court opinions from 2000 to 2004 and coding factor outcomes as favoring or disfavoring likelihood of confusion). The study's contributions were substantial but methodologically constrained: district court opinions may not represent the full population of trademark disputes, and binary coding cannot capture the degree to which factors influence outcomes. *See id.* at 1592–93.

[417] *See* Michael Heise, *The Importance of Being Empirical*, 26 PEPP. L. REV. 807, 819–24 (1999) (documenting the gap between empirical scholarship and judicial practice, and noting that doctrinal change typically lags behind empirical findings by years or decades).

[418] AM. INTELL. PROP. L. ASS'N, 2023 REPORT OF THE ECONOMIC SURVEY 57–63 (2023).

[419] Shari Seidman Diamond, *Reference Guide on Survey Research*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 359, 389–95 (Fed. Judicial Ctr. & Nat'l Acads. 3d ed. 2011); *What to Expect in*

generating evidence that may not matter, addressing factors that do not predict outcomes, and briefing issues whose resolution is foreordained by the real drivers of decision.

Beyond aggregate waste, these costs fall unevenly. Marc Galanter's foundational work identified the structural advantages "repeat players" enjoy over "one-shotters" in litigation.[420] Repeat players accumulate expertise, develop favorable precedent, and absorb individual losses within broader enforcement portfolios. Procedural complexity advantages them because they amortize learning costs across multiple matters.[421] The multifactor framework amplifies these asymmetries. A major brand holder with an in-house trademark team can litigate confusion factors efficiently; the startup facing its first opposition cannot. The framework's celebrated flexibility, in practice, favors those with resources to exploit it.

This dynamic manifests in what the USPTO has termed "trademark bullying": conduct where a trademark owner "uses its trademark rights to harass and intimidate another business beyond what the law might be reasonably interpreted to allow."[422] The USPTO's 2011 Report to Congress acknowledged that small businesses reported abandoning applications after receiving cease-and-desist letters from larger companies, not because the claims had merit but because respondents lacked resources to litigate.[423] As eBay observed in its comments: "Trademark bullies are successful in obtaining settlements against trademark owners, even where the trademark infringement claims asserted are questionable, because defending parties are either not capable, financially or otherwise, or willing to deal with the risks and uncertainty involved in litigating a trademark dispute."[424]

The framework enables this bullying by generating uncertainty. If outcomes turned predictably on two factors, parties could evaluate demands rationally: compare the marks, compare the goods, estimate the result. But when outcomes theoretically depend on thirteen factors weighted through discretionary balancing, prediction becomes difficult. This creates what Mnookin and Kornhauser identified as a "bargaining backdrop clouded by uncertainty."[425] Clear legal rules facilitate settlement by allowing parties to negotiate in the shadow of predictable outcomes. When the shadow grows murky, the resource-constrained party may capitulate regardless of merits.

2. The Legitimacy Deficit

The economic costs are measurable, if dispiriting. The institutional costs are harder to quantify but potentially more corrosive. When legal doctrine systematically diverges from legal practice, the gap undermines the system's claim to legitimacy.

Trademark Litigation, GERBEN IP, https://www.gerbenlaw.com/blog/what-to-expect-in-trademark-litigation/ (last visited Dec. 1, 2025) (reporting survey and expert costs of $150,000–$200,000).

[420] Marc Galanter, *Why the "Haves" Come Out Ahead: Speculations on the Limits of Legal Change*, 9 LAW & SOC'Y REV. 95, 97–104 (1974).

[421] *Id.* at 103-08, 114-19.

[422] U.S. PATENT & TRADEMARK OFFICE, TRADEMARK LITIGATION TACTICS AND FEDERAL GOVERNMENT SERVICES TO PROTECT TRADEMARKS AND PREVENT COUNTERFEITING: A REPORT TO CONGRESS 4 (Apr. 2011).

[423] *Id.* at 19-23.

[424] Comments of eBay Inc. to the USPTO (Dec. 17, 2010).

[425] Robert H. Mnookin & Lewis Kornhauser, *Bargaining in the Shadow of the Law: The Case of Divorce*, 88 YALE L.J. 950, 968–69 (1979).

Lon Fuller's classic account of law's "internal morality" identified eight principles that distinguish genuine legal systems from mere exercises of power.[426] Among these: laws must be publicly promulgated, sufficiently clear to guide conduct, and administered congruently with their declared content.[427] A regime that announces one set of rules while applying another fails Fuller's test. It may technically function, but it forfeits the moral authority that distinguishes law from force.

The DuPont framework falls uncomfortably close to this line. The doctrine announces a thirteen-factor test; practice applies two. The doctrine proclaims that no factor is dispositive; practice treats similarity and proximity as nearly so. The doctrine insists on case-by-case balancing; practice produces outcomes predictable from a fraction of the inputs. This is not quite the "secret law" Fuller condemned, but it approaches what we might call "insider law": rules whose actual operation is legible primarily to initiates.

The problem is not that trademark specialists understand how the system really works. Expertise always confers advantage. The problem is that the system's public face misrepresents its actual operation. A small business owner reading DuPont or its progeny would reasonably conclude that intent matters, that survey evidence is important, that the strength of the senior mark could prove decisive. She would be wrong on all counts, but she would be wrong in precisely the way the doctrine invited her to be. The framework does not merely fail to guide; it actively misleads.

Tom Tyler's research on procedural justice demonstrates why this matters beyond the individual case.[428] Public compliance with law depends less on fear of sanctions than on perceptions of legitimacy. People obey legal authorities they perceive as fair and trustworthy; they resist those they perceive as arbitrary or illegitimate.[429] Legitimacy, in turn, depends partly on whether the system operates as advertised. A legal regime that says one thing and does another invites the cynicism that corrodes voluntary compliance.

Trademark law may seem too specialized to implicate these concerns. Most citizens will never litigate a confusion claim. But the broader lesson holds: legal systems purchase compliance with coherence. When doctrine and practice diverge, the currency is debased. Practitioners learn to discount official pronouncements; parties learn to distrust predictability; observers learn that law is less a system of rules than a vocabulary for rationalizing preferred outcomes. None of this serves trademark law's legitimate functions.

It need not be so. Trademark law could align its doctrine with its practice, acknowledge what actually drives decisions, and offer parties a framework that means what it says.

[426] LON L. FULLER, THE MORALITY OF LAW 46–91 (rev. ed. 1969).

[427] *Id.* at 49–51 (promulgation), 63–65 (clarity), 81–91 (congruence between official action and declared rule).

[428] TOM R. TYLER, WHY PEOPLE OBEY THE LAW 57–87 (2006).

[429] *Id.* at 269–75 (discussing relationship between procedural fairness, legitimacy, and voluntary compliance).

PART IV

THE POST-MORTEM: WHERE TO GO FROM HERE

A. Research Agenda

This study provides the first large-scale computational analysis of TTAB likelihood-of-confusion decisions. The findings are robust: of thirteen DuPont factors, only two consistently predict outcomes. But TTAB decisions represent just one slice of trademark confusion jurisprudence. Several extensions would strengthen and refine these conclusions.

First, the methodology developed here should be applied to federal circuit court decisions. Preliminary analysis of Federal Circuit and regional circuit opinions suggests similar patterns, but the sample sizes remain small. A comprehensive study coding all published circuit court trademark opinions since DuPont would test whether appellate courts exhibit the same factor-outcome relationships observed at the TTAB, or whether appellate review introduces meaningful correction.

Second, international comparison would illuminate whether multifactor collapse reflects something inherent to confusion analysis or something peculiar to American doctrine. The European Union's likelihood-of-confusion test employs fewer factors with explicit weighting guidance.[430] The United Kingdom's approach differs again.[431] Comparative empirical analysis could identify whether alternative doctrinal structures produce more predictable or more accurate outcomes.

Third, the corpus itself invites continued development. Machine-learning classification of factor outcomes enables analysis at scale previously impossible. As new decisions issue, the model can be updated, permitting longitudinal tracking of doctrinal evolution. If the Federal Circuit eventually acknowledges the empirical reality documented here, the corpus would capture any resulting shift in TTAB practice.

These extensions matter, but they should not obscure what the present study has already established: the thirteen-factor test for trademark confusion does not function as advertised. Courts genuinely weigh two factors; eleven others serve as window dressing. This finding has implications beyond trademark law.

B. The Multifactor Collapse Hypothesis

The dysfunction documented in Parts II and III is not unique to trademark law. Across legal doctrine, courts employ elaborate multifactor tests that purport to weigh numerous considerations but actually turn on one or two variables. Call this the *multifactor collapse hypothesis*: when courts apply balancing tests with more than three or four factors, the test collapses in practice to a smaller number of determinative considerations, while the remaining factors serve rhetorical rather than decisional functions.

The hypothesis finds support in multiple domains. Consider copyright's four-factor fair use test.[432] Barton Beebe's empirical studies reveal that transformativeness dominates the

[430] *See* Council Regulation 2017/1001, art. 8(1)(b), 2017 O.J. (L 154) 1 (EU) (establishing likelihood of confusion standard without enumerated factors); ILANAH SIMON FHIMA, TRADE MARK DILUTION IN EUROPE AND THE UNITED STATES 50–58 (2011) (comparing European and American approaches to confusion analysis).

[431] *See* Trade Marks Act 1994, c. 26, § 10(2) (UK); *Specsavers Int'l Healthcare Ltd. v. Asda Stores Ltd.*, [2012] EWCA (Civ) 24, [87] (Eng.) (applying "global appreciation" test for likelihood of confusion).

[432] 17 U.S.C. § 107 (2018) (enumerating four factors: purpose and character of use, nature of copyrighted work, amount used, and market effect).

analysis: when courts find a use transformative, fair use follows roughly 92% of the time.[433] Beebe observed that transformativeness exerts "nearly dispositive force not simply on the outcome of factor one but on the overall outcome of the fair use test."[434] The statutory four-factor structure remains formally intact, but actual decision-making has collapsed into a single inquiry. Courts discuss all four factors, but the outcome is effectively determined before factors two through four receive consideration.

Qualified immunity offers another illustration. The doctrine requires plaintiffs to show that defendants violated "clearly established law," a standard the Supreme Court describes as "protecting all but the plainly incompetent or those who knowingly violate the law."[435] Commentators describe qualified immunity in stark terms, arguing it slams courthouse doors on meritorious claims. Yet Joanna Schwartz's empirical study of 1,183 Section 1983 cases found that qualified immunity caused dismissal in only 3.9% of cases.[436] The doctrine doesn't function as either supporters or critics describe. Cases that fail do so for other reasons; qualified immunity's elaborate "clearly established" analysis rarely disposes of litigation.[437] The doctrinal framework persists, but actual outcomes turn on different considerations entirely.

Personal jurisdiction doctrine exhibits a similar pattern. World-Wide Volkswagen articulated five reasonableness factors: burden on the defendant, forum state's interest, plaintiff's interest in convenient relief, the interstate judicial system's interest in efficient resolution, and shared interests in substantive social policies.[438] Yet contemporary courts have "relegated the fairness prong of this test to, at most, an afterthought."[439] Purposeful availment effectively determines jurisdiction; the five-factor reasonableness analysis persists in opinions but contributes little to outcomes.

These examples suggest a general phenomenon. Multifactor tests appeal to courts and legislatures because they signal comprehensiveness. Listing many factors creates an impression of careful balancing, cabining judicial discretion, and attending to contextual nuance.[440] But human cognition resists genuine multifactor balancing. Judges, like other decision-makers, rely on heuristics that prioritize the most salient and diagnostic variables.[441] Elaborate frameworks obscure rather than guide this process.

[433] Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions Updated, 1978–2019*, 10 N.Y.U. J. INTELL. PROP. & ENT. L. 1, 24–27 (2020) (finding 92.2% of cases with factor one favoring fair use resulted in fair use findings; odds ratio of 91.3:1 for transformative use leading to fair use).

[434] Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978–2005*, 156 U. PA. L. REV. 549, 605–06 (2008).

[435] *Malley v. Briggs*, 475 U.S. 335, 341 (1986).

[436] Joanna C. Schwartz, *How Qualified Immunity Fails*, 127 YALE L.J. 2, 45–47 (2017) (reviewing 1,183 Section 1983 cases and finding qualified immunity disposed of only 38 cases, or 3.9% of cases where the defense could be raised).

[437] *Id.* at 50–54 (explaining that qualified immunity rarely shields defendants from discovery and trial burdens as the doctrine intends).

[438] *World-Wide Volkswagen Corp. v. Woodson*, 444 U.S. 286, 292 (1980).

[439] Megan M. La Belle, *Personal Jurisdiction and the Fairness Factor(s)*, 72 EMORY L.J. 781, 785 (2023); *see also* A. Benjamin Spencer, *Jurisdiction to Adjudicate: A Revised Analysis*, 73 U. CHI. L. REV. 617, 623 (2006) ("The burden on defendants is typically given the most weight, with the plaintiffs' interests and state interests receiving a fair degree of consideration as well.").

[440] *See* Schauer, *Formalism*, *supra* note 399, at 539–44 (analyzing how formal legal structures signal comprehensiveness while channeling discretion).

[441] Guthrie, Rachlinski & Wistrich, *supra* note 394, at 784–88 (presenting empirical evidence that federal magistrate judges, despite expertise and motivation, remain susceptible to cognitive heuristics and biases).

Despite the imagery the term suggests, consider that *collapse is not necessarily bad*. If two factors genuinely determine outcomes, pretending otherwise wastes resources and obscures doctrine. Trademark law would be improved, not degraded, by acknowledging that mark similarity and goods relatedness drive confusion analysis. Honest doctrine would reduce litigation costs, improve settlement behavior, and enhance rule-of-law values by aligning what courts say with what courts do.

The normative case for simplification is straightforward: transparency, efficiency, and legitimacy. But doctrinal reform faces powerful obstacles. Stare decisis binds courts to existing frameworks. Practitioners have invested in learning current doctrine. Simplification requires acknowledging that prior judicial rhetoric overstated the comprehensiveness of analysis. These barriers explain why empirical findings rarely produce rapid doctrinal change.

This Article has provided definitive evidence that DuPont's thirteen factors do not function as advertised. Eleven factors are decorative. The comprehensive balancing framework is performance, not practice.

This finding should prompt humility. If a doctrine recited tens of thousands of times can operate so differently from its stated form, what else do we not know about how law actually works? Empirical legal scholarship has only begun to illuminate the gap between doctrine on the books and doctrine in action. DuPont is merely one data point. The legal system is replete with multifactor tests, totality-of-circumstances standards, and balancing frameworks that have never been rigorously examined. The tools now exist to examine them. The barrier is no longer feasibility, but fortitude: will we confront the law as it operates, or continue to recite it as we wish it did?